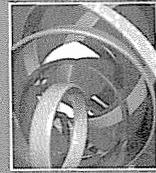


MICROECONOMICS

FIFTH EDITION

Robert S. Pindyck
Daniel L. Rubinfeld

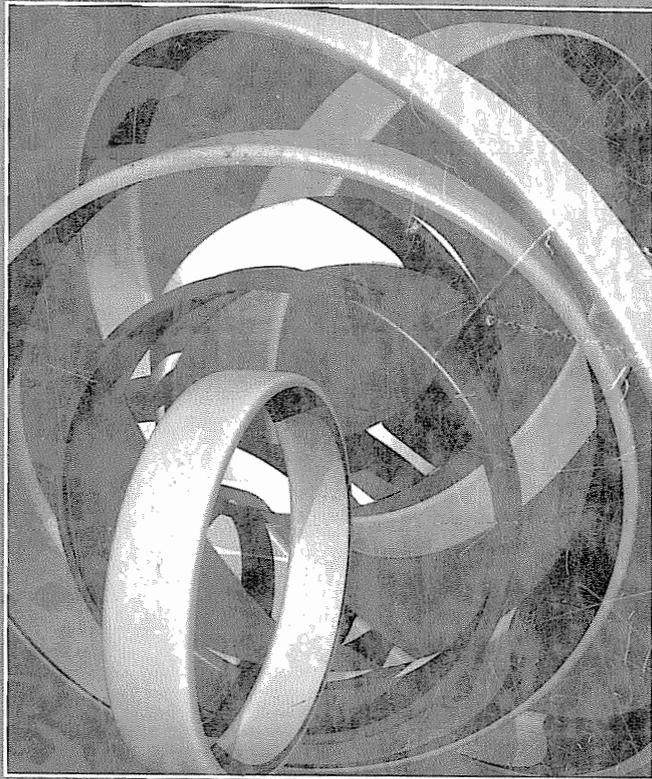
PINDYCK
RUBINFELD



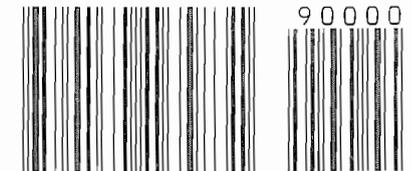
MICROECONOMICS

FIFTH
EDITION

Prentice
Hall



ISBN 0-13-016583-2



9 780130 165831

90000

Prentice Hall
Saddle River, New Jersey 07458
prenhall.com

R eal-world examples are key to the applied approach of this book. The fifth edition of *Microeconomics* incorporates more than 100 detailed examples into the flow of the text. The following is a list of these examples:

CHAPTER	EXAMPLE	TOPIC	PAGE REFERENCE
1	1.1	Markets for Prescription Drugs	10
	1.2	The Price of Eggs and the Price of a College Education	12
	1.3	The Minimum Wage	13
2	2.1	The Price of Eggs and Price of a College Education Revisited	26
	2.2	Wage Inequality in the United States	27
	2.3	The Long-Run Behavior of Natural Resource Prices	28
	2.4	The Market for Wheat	33
	2.5	The Demand for Gasoline and Automobiles	39
	2.6	The Weather in Brazil and the Price of Coffee in New York	41
	2.7	Declining Demand and the Behavior of Copper Prices	47
	2.8	Upheaval in the World Oil Market	49
	2.9	Price Controls and Natural Gas Shortages	54
3	3.1	Designing New Automobiles (I)	71
	3.2	Designing New Automobiles (II)	81
	3.3	Decision Making and Public Policy	82
	3.4	A College Trust Fund	85
	3.5	Revealed Preference for Recreation	88
	3.6	Gasoline Rationing	91
	3.7	The Bias in the CPI	97
4	4.1	Consumer Expenditures in the United States	108
	4.2	The Effects of a Gasoline Tax	114
	4.3	The Aggregate Demand for Wheat	120
	4.4	The Demand for Housing	122
	4.5	The Value of Clean Air	125
	4.6	Network Externalities and Demands for Computers and E-Mail	130
	4.7	The Demand for Ready-to-Eat Cereal	134
5	5.1	Deterring Crime	154
	5.2	Business Executives and the Choice of Risk	160
	5.3	The Value of Title Insurance When Buying a House	163
	5.4	The Value of Information in the Dairy Industry	165
	5.5	Investing in the Stock Market	173
6	6.1	Malthus and the Food Crisis	187
	6.2	Labor Productivity and the Standard of Living	189
	6.3	A Production Function for Wheat	196
	6.4	Returns to Scale in the Carpet Industry	199
7	7.1	Choosing the Location for a New Law School Building	205
	7.2	Sunk, Fixed, and Variable Costs: Computers, Software, and Pizzas	207
	7.3	The Short-Run Cost of Aluminum Smelting	213
	7.4	The Effect of Effluent Fees on Input Choices	220
	7.5	Economies of Scope in the Trucking Industry	232
	7.6	The Learning Curve in Practice	236
	7.7	Cost Functions for Electric Power	240
	7.8	A Cost Function for the Savings and Loan Industry	241
8	8.1	The Short-Run Output Decision of an Aluminum Smelting Plant	260
	8.2	Some Cost Considerations for Managers	261
	8.3	The Short-Run Production of Petroleum Products	265
	8.4	The Short-Run World Supply of Copper	268
	8.5	The Long-Run Supply of Housing	282
9	9.1	Price Controls and Natural Gas Shortages	292
	9.2	The Market for Human Kidneys	295
	9.3	Airline Regulation	300
	9.4	Supporting the Price of Wheat	306
	9.5	The Sugar Quota	312
	9.6	A Tax on Gasoline	318

CHAPTER	EXAMPLE	TOPIC	PAGE REFERENCE
10	10.1	Astra-Merck Prices Prilosec	334
	10.2	Markup Pricing: Supermarkets to Designer Jeans	342
	10.3	The Pricing of Pre-recorded Videocassettes	343
	10.4	Monopsony Power in U.S. Manufacturing	358
	10.5	A Phone Call About Prices	362
	10.6	The United States versus Microsoft	363
11	11.1	The Economics of Coupons and Rebates	379
	11.2	Airline Fares	380
	11.3	How to Price a Best-Selling Novel	384
	11.4	Polaroid Cameras	389
	11.5	Pricing Cellular Phone Service	390
	11.6	The Complete Dinner versus a la Carte: A Restaurant's Pricing Problem	401
	11.7	Advertising in Practice	406
12	12.1	Monopolistic Competition in the Markets for Colas and Coffee	428
	12.2	A Pricing Problem for Procter & Gamble	440
	12.3	Procter & Gamble in a Prisoners' Dilemma	444
	12.4	Price Leadership and Price Rigidity in Commercial Banking	448
	12.5	The Cartelization of Intercollegiate Athletics	455
	12.6	The Milk Cartel	456
13	13.1	Acquiring a Company	463
	13.2	Oligopolistic Cooperation in the Water Meter Industry	474
	13.3	Competition and Collusion in the Airline Industry	475
	13.4	Wal-Mart Stores' Preemptive Investment Strategy	482
	13.5	DuPont Deters Entry in the Titanium Dioxide Industry	487
	13.6	Diaper Wars	488
	13.7	Internet Auctions	495
14	14.1	The Demand for Jet Fuel	508
	14.2	Labor Supply for One- and Two-Earner Households	513
	14.3	Pay in the Military	517
	14.4	Monopsony Power in the Market for Baseball Players	520
	14.5	Teenage Labor Markets and the Minimum Wage	521
	14.6	The Decline of Private-Sector Unionism	527
	14.7	Wage Inequality—Have Computers Changed the Labor Market?	528
15	15.1	The Value of Lost Earnings	537
	15.2	The Yields on Corporate Bonds	541
	15.3	Capital Investment in the Disposable Diaper Industry	548
	15.4	Choosing an Air Conditioner and a New Car	550
	15.5	How Depletable Are Depletable Resources?	554
16	16.1	The Interdependence of International Markets	566
	16.2	The Effects of Automobile Import Quotas	588
	16.3	The Costs and Benefits of Special Protection	589
	16.4	Lemons in Major League Baseball	600
	16.5	Working into the Night	605
17	17.1	Reducing Moral Hazard—Warranties of Animal Health	608
	17.2	Crisis in the Savings and Loan Industry	608
	17.3	Managers of Nonprofit Hospitals as Agents	611
	17.4	Efficiency Wages at Ford Motor Company	618
	17.5	The Costs and Benefits of Reduced Sulfur Dioxide Emissions	631
	17.6	Emissions Trading and Clean Air	632
	17.7	Regulating Municipal Solid Wastes	637
	17.8	The Coase Theorem at Work	641
18	18.1	Crawfish Fishing in Louisiana	643
	18.2	The Demand for Clean Air	647
	18.3		
	18.4		

PRENTICE HALL SERIES IN ECONOMICS

- Adams/Brock, *The Structure of American Industry*, Tenth Edition
- Blanchard, *Macroeconomics*, Second Edition
- Blau/Ferber/Winkler, *The Economics of Women, Men, and Work*, Third Edition
- Boardman/Greenberg/Vining/Wiemer, *Cost Benefit Analysis: Concepts and Practice*, Second Edition
- Bogart, *The Economics of Cities and Suburbs*
- Case/Fair, *Principles of Economics*, Fifth Edition
- Case/Fair, *Principles of Macroeconomics*, Fifth Edition
- Case/Fair, *Principles of Microeconomics*, Fifth Edition
- Caves, *American Industry: Structure, Conduct, Performance*, Seventh Edition
- Collinge/Ayers, *Economics by Design: Principles and Issues*, Second Edition
- DiPasquale/Wheaton, *Urban Economics and Real Estate Markets*
- Feiner, *Race and Gender in the American Economy: Views Across the Spectrum*
- Folland/Goodman/Stano, *Economics of Health and Health Care*, Third Edition
- Froyen, *Macroeconomics: Theories and Policies*, Sixth Edition
- Greene, *Econometric Analysis*, Fourth Edition
- Heilbroner/Milberg, *The Making of an Economic Society*, Tenth Edition
- Heyne, *The Economic Way of Thinking*, Ninth Edition
- Hirschleifer/Hirschleifer, *Price Theory and Applications*, Sixth Edition
- Keat/Young, *Managerial Economics*, Third Edition
- Milgrom/Roberts, *Economics, Organization, and Management*
- O'Sullivan/Sheffrin, *Economics: Principles and Tools*, Second Edition
- O'Sullivan/Sheffrin, *Macroeconomics: Principles and Tools*, Second Edition
- O'Sullivan/Sheffrin, *Microeconomics: Principles and Tools*, Second Edition
- Petersen/Lewis, *Managerial Economics*, Fourth Edition
- Pindyck/Rubinfeld, *Microeconomics*, Fifth Edition
- Reynolds/Masters/Moser, *Labor Economics and Labor Relations*, Eleventh Edition
- Roberts, *The Choice: A Fable of Free Trade and Protectionism*, Revised
- Sachs/Larrain, *Macroeconomics in the Global Economy*
- Schiller, *The Economics of Poverty and Discrimination*, Eighth Edition
- Weidenbaum, *Business and Government in the Global Marketplace*, Sixth Edition

MICROECONOMICS

FIFTH EDITION

Robert S. Pindyck
Massachusetts Institute of Technology

Daniel L. Rubinfeld
University of California, Berkeley



Prentice Hall, Upper Saddle River, New Jersey 07458

Library of Congress Cataloging-in-Publication Data

Pindyck, Robert S.

Microeconomics/Robert S. Pindyck, Daniel L. Rubinfeld.—

5th ed.

p. cm. — (Prentice-Hall series in economics)

Multi-media teaching aids available to supplement the text.

Includes bibliographical references and index.

ISBN 0-13-016583-2

1. Microeconomics. I. Rubinfeld, Daniel L. II. Title. III. Series.

HB172.P53 2000

338.5—dc21

00-035995

Senior Editor: Rod Banister

Managing Editor (Editorial): Gladys Soto

Editor-in-Chief: PJ Boardman

Editorial Assistant: Marie McHale

Assistant Editor: Holly Brown

Media Project Manager: Bill Minick

Senior Marketing Manager: Lori Braumberger

Managing Editor (Production): Cynthia Regan

Production Coordinator: Elena Barnett

Manufacturing Supervisor: Paul Smolenski

Manufacturing Buyer: Lisa Babin

Senior Prepress/Manufacturing Manager: Vincent Scelta

Design Manager: Patricia Smythe

Cover/Interior Design: Lorraine Castellano

Cover Art: Steven Gagliostro

Composition: York Graphic Services, Inc.

Project Management: York Production Services



To our daughters,

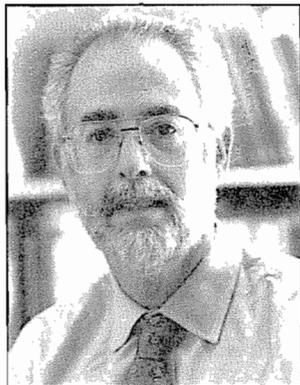
*Maya, Talia, and Shira
Sarah and Rachel*

Copyright © 2001, 1998, 1995 by Prentice-Hall, Inc., Upper Saddle River, New Jersey, 07458. All rights reserved. Printed in the United States of America. This publication is protected by Copyright and permission should be obtained from the publisher prior to any prohibited reproduction, storage in a retrieval system, or transmission in any form or by any means, electronic, mechanical, photocopying, recording, or likewise. For information regarding permission(s), write to: Rights and Permissions Department.

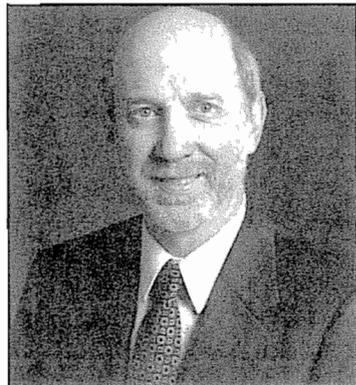


10 9 8 7 6 5 4 3
ISBN 0-13-016583-2

ABOUT THE AUTHORS



Professor Robert S. Pindyck



Professor Daniel L. Rubinfeld

Robert S. Pindyck is the Mitsubishi Bank Professor in Economics and Finance in the Sloan School of Management at M.I.T. He is also a Research Associate of the National Bureau of Economic Research, and a Fellow of the Econometric Society, and has been a Visiting Professor of Economics at Tel-Aviv University. He received his Ph.D. in Economics from M.I.T. in 1971. Professor Pindyck's research and writing have covered a variety of topics in microeconomics and industrial organization, including the effects of uncertainty on firm behavior and market structure, determinants of market power, the behavior of natural resource, commodity, and financial markets, and criteria for investment decisions. He has been a consultant to a number of public and private organizations, and is currently co-editor of *The Review of Economics and Statistics*. He is also the co-author with Daniel Rubinfeld of *Econometric Models and Economic Forecasts*, a best-selling textbook that may or may not be turned into a feature film.

Daniel L. Rubinfeld is Robert L. Bridges Professor of Law and Professor of Economics at the University of California, Berkeley. He taught previously at Suffolk University, Wellesley College, and the University of Michigan, and served from June 1997 through December 1998 as Deputy Assistant Attorney General for Antitrust in the U.S. Department of Justice. He has been a Fellow at the National Bureau of Economic Research, the Center for Advanced Study in the Behavioral Sciences, and the John Simon Guggenheim Foundation. He received a BA in mathematics from Princeton University in 1967 and a Ph.D. in Economics from M.I.T. in 1972. Professor Rubinfeld is the author of a variety of articles relating to competition policy, law and economics, law and statistics, and public economics. He is currently co-editor of the *International Review of Law and Economics*, and has served as Associate Dean and Chair of the Jurisprudence and Social Policy Program at Berkeley from 1987–1990 and 1999–2000. He is the co-author (with Robert Pindyck) of *Econometric Models and Economic Forecasts*, and expects to play the lead in the film version of the book.

BRIEF CONTENTS

PART 1	Introduction: Markets and Prices 1
	1 Preliminaries 3
	2 The Basics of Supply and Demand 19
PART 2	Producers, Consumers, and Competitive Markets 59
	3 Consumer Behavior 61
	4 Individual and Market Demand 101
	5 Choice Under Uncertainty 149
	6 Production 177
	7 The Cost of Production 203
	8 Profit Maximization and Competitive Supply 251
	9 The Analysis of Competitive Markets 287
PART 3	Market Structure and Competitive Strategy 325
	10 Market Power: Monopoly and Monopsony 327
	11 Pricing with Market Power 369
	12 Monopolistic Competition and Oligopoly 423
	13 Game Theory and Competitive Strategy 461
	14 Markets for Factor Inputs 501
	15 Investment, Time, and Capital Markets 533
PART 4	Information, Market Failure, and the Role of Government 561
	16 General Equilibrium and Economic Efficiency 563
	17 Markets with Asymmetric Information 595
	18 Externalities and Public Goods 621
	Appendix: The Basics of Regression 655
	Glossary 663
	Answers to Selected Exercises 675
	Index 687

CONTENTS

Preface xxiii

PART 1

Introduction: Markets and Prices 1

1 Preliminaries 3

- 1.1 The Themes of Microeconomics 4
 - Theories and Models* 5
 - Positive versus Normative Analysis* 6
- 1.2 What Is a Market? 7
 - Competitive versus Noncompetitive Markets* 8
 - Market Price* 8
 - Market Definition—The Extent of a Market* 9
- 1.3 Real versus Nominal Prices 11
- 1.4 Why Study Microeconomics? 15
 - Corporate Decision Making: Ford's Sport Utility Vehicles* 15
 - Public Policy Design: Automobile Emission Standards for the Twenty-first Century* 16
- Summary 17
- Questions for Review 17
- Exercises 18

2 The Basics of Supply and Demand 19

- 2.1 Supply and Demand 20
 - The Supply Curve* 20
 - The Demand Curve* 21
- 2.2 The Market Mechanism 23
- 2.3 Changes in Market Equilibrium 24
- 2.4 Elasticities of Supply and Demand 30
- 2.5 Short-Run versus Long-Run Elasticities 35
 - Demand* 35
 - Supply* 40
- *2.6 Understanding and Predicting the Effects of Changing Market Conditions 44
- 2.7 Effects of Government Intervention—Price Controls 53
- Summary 55
- Questions for Review 56
- Exercises 57

PART 2 Producers, Consumers, and Competitive Markets 59**3 Consumer Behavior 61**

- Consumer Behavior 61
- 3.1 Consumer Preferences 62
 - Market Baskets 62
 - Some Basic Assumptions About Preferences 63
 - Indifference Curves 64
 - Indifference Maps 66
 - The Shapes of Indifference Curves 67
 - The Marginal Rate of Substitution 68
 - Perfect Substitutes and Perfect Complements 69
- 3.2 Budget Constraints 75
 - The Budget Line 75
 - The Effects of Changes in Income and Prices 77
- 3.3 Consumer Choice 79
 - Corner Solutions 84
- 3.4 Revealed Preference 86
- 3.5 Marginal Utility and Consumer Choice 89
- *3.6 Cost-of-Living Indexes 92
 - Ideal Cost-of-Living Index 93
 - Laspeyres Index 94
 - Paasche Index 95
 - Chain-Weighted Indexes 96
- Summary 98
- Questions for Review 99
- Exercises 99

4 Individual and Market Demand 101

- 4.1 Individual Demand 102
 - Price Changes 102
 - The Individual Demand Curve 102
 - Income Changes 104
 - Normal versus Inferior Goods 106
 - Engel Curves 106
 - Substitutes and Complements 109
- 4.2 Income and Substitution Effects 110
 - Substitution Effect 111
 - Income Effect 112
 - A Special Case: The Giffen Good 113
- 4.3 Market Demand 116
 - From Individual to Market Demand 116
 - Elasticity of Demand 117
- 4.4 Consumer Surplus 123
 - Consumer Surplus and Demand 123
- 4.5 Network Externalities 127
 - The Bandwagon Effect 127
 - The Snob Effect 129

- *4.6 Empirical Estimation of Demand 131
 - Interview and Experimental Approaches to Demand Determination 131
 - The Statistical Approach to Demand Estimation 132
 - The Form of the Demand Relationship 133

Summary 135

Questions for Review 136

Exercises 136

Appendix to Chapter 4: Demand Theory—A Mathematical Treatment 139

- Utility Maximization 139
- The Method of Lagrange Multipliers 140
- The Equal Marginal Principle 141
- Marginal Rate of Substitution 141
- Marginal Utility of Income 142
- An Example 143
- Duality in Consumer Theory 144
- Income and Substitution Effects 145
- Exercises 147

5 Choice Under Uncertainty 149

- 5.1 Describing Risk 150
 - Probability 150
 - Expected Value 150
 - Variability 151
 - Decision Making 153
- 5.2 Preferences Toward Risk 155
 - Different Preferences Toward Risk 157
- 5.3 Reducing Risk 161
 - Diversification 161
 - Insurance 162
 - The Value of Information 164
- *5.4 The Demand for Risky Assets 166
 - Assets 166
 - Risky and Riskless Assets 166
 - Asset Returns 167
 - The Trade-Off Between Risk and Return 168
 - The Investor's Choice Problem 169

Summary 174

Questions for Review 175

Exercises 175

6 Production 177

- 6.1 The Technology of Production 178
 - The Production Function 178
- 6.2 Isoquants 179
 - Input Flexibility 180
 - The Short Run versus the Long Run 180
- 6.3 Production with One Variable Input (Labor) 181
 - Average and Marginal Products 182

	<i>The Slopes of the Product Curve</i>	183
	<i>The Average Product of Labor Curve</i>	184
	<i>The Marginal Product of Labor Curve</i>	185
	<i>The Law of Diminishing Marginal Returns</i>	185
	<i>Labor Productivity</i>	188
6.4	Production with Two Variable Inputs	191
	<i>Diminishing Marginal Returns</i>	191
	<i>Substitution Among Inputs</i>	192
	<i>Production Functions—Two Special Cases</i>	194
6.5	Returns to Scale	197
	<i>Increasing Returns to Scale</i>	198
	<i>Constant Returns to Scale</i>	198
	<i>Decreasing Returns to Scale</i>	198
	<i>Describing Returns to Scale</i>	198
	Summary	201
	Questions for Review	201
	Exercises	202
7	The Cost of Production	203
7.1	Measuring Cost: Which Costs Matter?	203
	<i>Economic Cost versus Accounting Cost</i>	204
	<i>Opportunity Cost</i>	204
	<i>Sunk Costs</i>	205
	<i>Fixed Costs and Variable Costs</i>	206
	<i>Fixed versus Sunk Costs</i>	207
7.2	Cost in the Short Run	208
	<i>The Determinants of Short-Run Cost</i>	210
	<i>The Shapes of the Cost Curves</i>	211
7.3	Cost in the Long Run	215
	<i>The User Cost of Capital</i>	215
	<i>The Cost-Minimizing Input Choice</i>	216
	<i>The Isocost Line</i>	217
	<i>Choosing Inputs</i>	218
	<i>Cost Minimization with Varying Output Levels</i>	222
	<i>The Expansion Path and Long-Run Costs</i>	222
7.4	Long-Run versus Short-Run Cost Curves	224
	<i>The Inflexibility of Short-Run Production</i>	224
	<i>Long-Run Average Cost</i>	225
	<i>Economies and Diseconomies of Scale</i>	227
	<i>The Relationship Between Short-Run and Long-Run Cost</i>	227
7.5	Production with Two Outputs—Economies of Scope	229
	<i>Product Transformation Curves</i>	230
	<i>Economies and Diseconomies of Scope</i>	231
	<i>The Degree of Economies of Scope</i>	231
*7.6	Dynamic Changes in Costs—The Learning Curve	232
	<i>Graphing the Learning Curve</i>	234
	<i>Learning versus Economies of Scale</i>	234

*7.7	Estimating and Predicting Cost	237
	<i>Cost Functions and the Measurement of Scale Economies</i>	239
	Summary	242
	Questions for Review	243
	Exercises	243

Appendix to Chapter 7: Production and Cost Theory—

A	Mathematical Treatment	246
	<i>Cost Minimization</i>	246
	<i>Marginal Rate of Technical Substitution</i>	247
	<i>Duality in Production and Cost Theory</i>	248
	<i>The Cobb-Douglas Cost and Production Functions</i>	248
	Exercises	250

8 Profit Maximization and Competitive Supply 251

8.1	Perfectly Competitive Markets	252
	<i>When Is a Market Highly Competitive?</i>	253
8.2	Profit Maximization	254
	<i>Do Firms Maximize Profit?</i>	254
8.3	Marginal Revenue, Marginal Cost, and Profit Maximization	255
	<i>Demand and Marginal Revenue for a Competitive Firm</i>	256
	<i>Profit Maximization by a Competitive Firm</i>	257
8.4	Choosing Output in the Short Run	258
	<i>Short-Run Profit Maximization by a Competitive Firm</i>	258
	<i>The Short-Run Profit of a Competitive Firm</i>	259
8.5	The Competitive Firm's Short-Run Supply Curve	263
	<i>The Firm's Response to an Input Price Change</i>	264
8.6	The Short-Run Market Supply Curve	266
	<i>Elasticity of Market Supply</i>	266
	<i>Producer Surplus in the Short Run</i>	269
8.7	Choosing Output in the Long Run	271
	<i>Long-Run Profit Maximization</i>	271
	<i>Long-Run Competitive Equilibrium</i>	272
	<i>Economic Rent</i>	275
	<i>Producer Surplus in the Long Run</i>	276
8.8	The Industry's Long-Run Supply Curve	277
	<i>Constant-Cost Industry</i>	277
	<i>Increasing-Cost Industry</i>	279
	<i>Decreasing-Cost Industry</i>	280
	<i>The Effects of a Tax</i>	280
	<i>Long-Run Elasticity of Supply</i>	281
	Summary	283
	Questions for Review	284
	Exercises	284

9 The Analysis of Competitive Markets 287

9.1	Evaluating the Gains and Losses from Government Policies—	
	<i>Consumer and Producer Surplus</i>	288
	<i>Review of Consumer and Producer Surplus</i>	288
	<i>Application of Consumer and Producer Surplus</i>	289

- 9.2 The Efficiency of a Competitive Market 294
- 9.3 Minimum Prices 298
- 9.4 Price Supports and Production Quotas 302
 - Price Supports* 302
 - Production Quotas* 304
- 9.5 Import Quotas and Tariffs 309
- 9.6 The Impact of a Tax or Subsidy 313
 - The Effects of a Subsidy* 317
- Summary 320
- Questions for Review 320
- Exercises 321

PART 3**Market Structure and Competitive Strategy 325****10 Market Power: Monopoly and Monopsony 327**

- 10.1 Monopoly 328
 - Average Revenue and Marginal Revenue* 328
 - The Monopolist's Output Decision* 329
 - An Example* 331
 - A Rule of Thumb for Pricing* 333
 - Shifts in Demand* 335
 - The Effect of a Tax* 335
 - The Multiplant Firm* 337
- 10.2 Monopoly Power 339
 - Measuring Monopoly Power* 340
 - The Rule of Thumb for Pricing* 341
- 10.3 Sources of Monopoly Power 345
 - The Elasticity of Market Demand* 345
 - The Number of Firms* 345
 - The Interaction Among Firms* 346
- 10.4 The Social Costs of Monopoly Power 347
 - Rent Seeking* 348
 - Price Regulation* 348
 - Natural Monopoly* 350
 - Regulations in Practice* 351
- 10.5 Monopsony 352
 - Monopsony and Monopoly Compared* 354
- 10.6 Monopsony Power 355
 - Sources of Monopsony Power* 356
 - The Social Costs of Monopsony Power* 357
 - Bilateral Monopoly* 358
- 10.7 Limiting Market Power: The Antitrust Laws 359
 - Enforcement of the Antitrust Laws* 361
- Summary 364
- Questions for Review 365
- Exercises 365

11 Pricing with Market Power 369

- 11.1 Capturing Consumer Surplus 370
- 11.2 Price Discrimination 371
 - First-Degree Price Discrimination* 371
 - Second-Degree Price Discrimination* 374
 - Third-Degree Price Discrimination* 375
- 11.3 Intertemporal Price Discrimination and Peak-Load Pricing 382
 - Intertemporal Price Discrimination* 382
 - Peak-Load Pricing* 383
- 11.4 The Two-Part Tariff 385
- *11.5 Bundling 392
 - Relative Valuations* 393
 - Mixed Bundling* 397
 - Bundling in Practice* 399
 - Tying* 402
- *11.6 Advertising 403
 - A Rule of Thumb for Advertising* 405
- Summary 407
- Questions for Review 408
- Exercises 408

Appendix to Chapter 11: Transfer Pricing in the Integrated Firm 413

- Transfer Pricing When There Is No Outside Market* 413
- Transfer Pricing with a Competitive Outside Market* 415
- Transfer Pricing with a Noncompetitive Outside Market* 417
- A Numerical Example* 420
- Exercises* 421

12 Monopolistic Competition and Oligopoly 423

- 12.1 Monopolistic Competition 424
 - The Makings of Monopolistic Competition* 424
 - Equilibrium in the Short Run and the Long Run* 425
 - Monopolistic Competition and Economic Efficiency* 426
- 12.2 Oligopoly 429
 - Equilibrium in an Oligopolistic Market* 430
 - The Cournot Model* 431
 - The Linear Demand Curve—An Example* 433
 - First Mover Advantage—The Stackelberg Model* 436
- 12.3 Price Competition 437
 - Price Competition with Homogeneous Products—The Bertrand Model* 437
 - Price Competition with Differentiated Products* 438
- 12.4 Competition versus Collusion: The Prisoners' Dilemma 442
- 12.5 Implications of the Prisoners' Dilemma for Oligopolistic Pricing 445
 - Price Rigidity* 446
 - Price Signaling and Price Leadership* 447
 - The Dominant Firm Model* 450
- 12.6 Cartels 451
 - Analysis of Cartel Pricing* 452

Summary 456
 Questions for Review 457
 Exercises 457

13 Game Theory and Competitive Strategy 461

13.1 Gaming and Strategic Decisions 461
Noncooperative versus Cooperative Games 462
 13.2 Dominant Strategies 464
 13.3 The Nash Equilibrium Revisited 466
Maximin Strategies 468
**Mixed Strategies* 470
 13.4 Repeated Games 472
 13.5 Sequential Games 476
The Extensive Form of a Game 477
The Advantage of Moving First 478
 13.6 Threats, Commitments, and Credibility 479
Empty Threats 480
Commitment and Credibility 480
 13.7 Entry Deterrence 483
Strategic Trade Policy and International Competition 485
 13.8 Bargaining Strategy 489
 *13.9 Auctions 491
Auction Formats 491
Valuation and Information 492
Private-Value Auctions 492
Common-Value Auctions 494
Maximizing Auction Revenue 495
 Summary 496
 Questions for Review 497
 Exercises 498

14 Markets for Factor Inputs 501

14.1 Competitive Factor Markets 501
Demand for a Factor Input When Only One Input Is Variable 502
Demand for a Factor Input When Several Inputs Are Variable 505
The Market Demand Curve 506
The Supply of Inputs to a Firm 509
The Market Supply of Inputs 511
 14.2 Equilibrium in a Competitive Factor Market 514
Economic Rent 515
 14.3 Factor Markets with Monopsony Power 518
Marginal and Average Expenditure 519
The Input Purchasing Decision of the Firm 520
 14.4 Factor Markets with Monopoly Power 523
Monopoly Power over the Wage Rate 523
Unionized and Nonunionized Workers 524
Bilateral Monopoly in the Labor Market 525

Summary 529
 Questions for Review 530
 Exercises 530

15 Investment, Time, and Capital Markets 533

15.1 Stocks versus Flows 534
 15.2 Present Discounted Value 534
Valuing Payment Streams 535
 15.3 The Value of a Bond 538
Perpetuities 538
The Effective Yield on a Bond 539
 15.4 The Net Present Value Criterion for Capital Investment Decisions 542
The Electric Motor Factory 543
Real versus Nominal Discount Rates 543
Negative Future Cash Flows 545
 15.5 Adjustments for Risk 545
Diversifiable versus Nondiversifiable Risk 546
The Capital Asset Pricing Model 547
 15.6 Investment Decisions by Consumers 549
 *15.7 Intertemporal Production Decisions—Depletable Resources 551
The Production Decision of an Individual Resource Producer 552
The Behavior of Market Price 553
User Cost 553
Resource Production by a Monopolist 554
 15.8 How Are Interest Rates Determined? 555
A Variety of Interest Rates 557
 Summary 558
 Questions for Review 558
 Exercises 559

PART 4

Information, Market Failure, and the Role of Government 561

16 General Equilibrium and Economic Efficiency 563

16.1 General Equilibrium Analysis 563
Two Interdependent Markets—Moving to General Equilibrium 564
The Attainment of General Equilibrium 565
 16.2 Efficiency in Exchange 567
The Advantages of Trade 568
The Edgeworth Box Diagram 569
Efficient Allocations 570
The Contract Curve 571
Consumer Equilibrium in a Competitive Market 572
The Economic Efficiency of Competitive Markets 574
 16.3 Equity and Efficiency 575
The Utility Possibilities Frontier 575
Equity and Perfect Competition 577

16.4	Efficiency in Production	578
	<i>Production in the Edgeworth Box</i>	578
	<i>Input Efficiency</i>	579
	<i>Producer Equilibrium in a Competitive Input Market</i>	580
	<i>The Production Possibilities Frontier</i>	581
	<i>Output Efficiency</i>	583
	<i>Efficiency in Output Markets</i>	584
16.5	The Gains from Free Trade	585
	<i>Comparative Advantage</i>	585
	<i>An Expanded Production Possibilities Frontier</i>	587
16.6	An Overview—The Efficiency of Competitive Markets	590
16.7	Why Markets Fail	591
	<i>Market Power</i>	592
	<i>Incomplete Information</i>	592
	<i>Externalities</i>	592
	<i>Public Goods</i>	593
	Summary	593
	Questions for Review	594
	Exercises	594
17	Markets with Asymmetric Information	595
17.1	Quality Uncertainty and the Market for Lemons	596
	<i>The Market for Used Cars</i>	596
	<i>Implications of Asymmetric Information</i>	598
	<i>The Importance of Reputation and Standardization</i>	599
17.2	Market Signaling	601
	<i>A Simple Model of Job Market Signaling</i>	602
	<i>Guarantees and Warranties</i>	604
17.3	Moral Hazard	606
17.4	The Principal–Agent Problem	609
	<i>The Principal–Agent Problem in Private Enterprises</i>	610
	<i>The Principal–Agent Problem in Public Enterprises</i>	610
	<i>Incentives in the Principal–Agent Framework</i>	612
*17.5	Managerial Incentives in an Integrated Firm	613
	<i>Asymmetric Information and Incentive Design in the Integrated Firm</i>	614
	<i>Applications</i>	616
17.6	Asymmetric Information in Labor Markets: Efficiency Wage Theory	616
	Summary	619
	Questions for Review	619
	Exercises	619
18	Externalities and Public Goods	621
18.1	Externalities	621
	<i>Negative Externalities and Inefficiency</i>	622
	<i>Positive Externalities and Inefficiency</i>	623
18.2	Ways of Correcting Market Failure	625
	<i>An Emissions Standard</i>	626
	<i>An Emissions Fee</i>	626

	<i>Standards versus Fees</i>	627
	<i>Transferable Emissions Permits</i>	630
	<i>Recycling</i>	634
18.3	Externalities and Property Rights	638
	<i>Property Rights</i>	638
	<i>Bargaining and Economic Efficiency</i>	638
	<i>Costly Bargaining—The Role of Strategic Behavior</i>	640
	<i>A Legal Solution—Suing for Damages</i>	640
18.4	Common Property Resources	642
18.5	Public Goods	644
	<i>Efficiency and Public Goods</i>	646
	<i>Public Goods and Market Failure</i>	647
18.6	Private Preferences for Public Goods	649
	Summary	651
	Questions for Review	651
	Exercises	652

APPENDIX

The Basics of Regression 655

	An Example	655
	Estimation	656
	Statistical Tests	657
	Goodness of Fit	659
	Economic Forecasting	660

Glossary 663

Answers to Selected Exercises 675

Index 687

List of Examples

Example 1.1	Markets for Prescription Drugs	10
Example 1.2	The Price of Eggs and the Price of a College Education	12
Example 1.3	The Minimum Wage	13
Example 2.1	The Price of Eggs and the Price of a College Education Revisited	26
Example 2.2	Wage Inequality in the United States	27
Example 2.3	The Long-Run Behavior of Natural Resource Prices	28
Example 2.4	The Market for Wheat	33
Example 2.5	The Demand for Gasoline and Automobiles	39
Example 2.6	The Weather in Brazil and the Price of Coffee in New York	41
Example 2.7	Declining Demand and the Behavior of Copper Prices	47
Example 2.8	Upheaval in the World Oil Market	49
Example 2.9	Price Controls and Natural Gas Shortages	54
Example 3.1	Designing New Automobiles (I)	71
Example 3.2	Designing New Automobiles (II)	81
Example 3.3	Decision Making and Public Policy	82
Example 3.4	A College Trust Fund	85

Example 3.5	Revealed Preference for Recreation	88
Example 3.6	Gasoline Rationing	91
Example 3.7	The Bias in the CPI	97
Example 4.1	Consumer Expenditures in the United States	108
Example 4.2	The Effects of a Gasoline Tax	114
Example 4.3	The Aggregate Demand for Wheat	120
Example 4.4	The Demand for Housing	122
Example 4.5	The Value of Clean Air	125
Example 4.6	Network Externalities and the Demands for Computers and E-Mail	130
Example 4.7	The Demand for Ready-to-Eat Cereal	134
Example 5.1	Deterring Crime	154
Example 5.2	Business Executives and the Choice of Risk	160
Example 5.3	The Value of Title Insurance When Buying a House	163
Example 5.4	The Value of Information in the Dairy Industry	165
Example 5.5	Investing in the Stock Market	173
Example 6.1	Malthus and the Food Crisis	187
Example 6.2	Labor Productivity and the Standard of Living	189
Example 6.3	A Production Function for Wheat	196
Example 6.4	Returns to Scale in the Carpet Industry	199
Example 7.1	Choosing the Location for a New Law School Building	205
Example 7.2	Sunk, Fixed, and Variable Costs: Computers, Software, and Pizzas	207
Example 7.3	The Short-Run Cost of Aluminum Smelting	213
Example 7.4	The Effect of Effluent Fees on Input Choices	220
Example 7.5	Economies of Scope in the Trucking Industry	232
Example 7.6	The Learning Curve in Practice	236
Example 7.7	Cost Functions for Electric Power	240
Example 7.8	A Cost Function for the Savings and Loan Industry	241
Example 8.1	The Short-Run Output Decision of an Aluminum Smelting Plant	260
Example 8.2	Some Cost Considerations for Managers	261
Example 8.3	The Short-Run Production of Petroleum Products	265
Example 8.4	The Short-Run World Supply of Copper	268
Example 8.5	The Long-Run Supply of Housing	282
Example 9.1	Price Controls and Natural Gas Shortages	292
Example 9.2	The Market for Human Kidneys	295
Example 9.3	Airline Regulation	300
Example 9.4	Supporting the Price of Wheat	306
Example 9.5	The Sugar Quota	312
Example 9.6	A Tax on Gasoline	318
Example 10.1	Astra-Merck Prices Prilosec	334
Example 10.2	Markup Pricing: Supermarkets to Designer Jeans	342
Example 10.3	The Pricing of Prerecorded Videocassettes	343

Example 10.4	Monopsony Power in U.S. Manufacturing	358
Example 10.5	A Phone Call About Prices	362
Example 10.6	The United States versus Microsoft	363
Example 11.1	The Economics of Coupons and Rebates	379
Example 11.2	Airline Fares	380
Example 11.3	How to Price a Best-Selling Novel	384
Example 11.4	Polaroid Cameras	389
Example 11.5	Pricing Cellular Phone Service	390
Example 11.6	The Complete Dinner versus à la Carte: A Restaurant's Pricing Problem	401
Example 11.7	Advertising in Practice	406
Example 12.1	Monopolistic Competition in the Markets for Colas and Coffee	428
Example 12.2	A Pricing Problem for Procter & Gamble	440
Example 12.3	Procter & Gamble in a Prisoners' Dilemma	444
Example 12.4	Price Leadership and Price Rigidity in Commercial Banking	448
Example 12.5	The Cartelization of Intercollegiate Athletics	455
Example 12.6	The Milk Cartel	456
Example 13.1	Acquiring a Company	463
Example 13.2	Oligopolistic Cooperation in the Water Meter Industry	474
Example 13.3	Competition and Collusion in the Airline Industry	475
Example 13.4	Wal-Mart Stores' Preemptive Investment Strategy	482
Example 13.5	DuPont Deters Entry in the Titanium Dioxide Industry	487
Example 13.6	Diaper Wars	488
Example 13.7	Internet Auctions	495
Example 14.1	The Demand for Jet Fuel	508
Example 14.2	Labor Supply for One- and Two-Earner Households	513
Example 14.3	Pay in the Military	517
Example 14.4	Monopsony Power in the Market for Baseball Players	520
Example 14.5	Teenage Labor Markets and the Minimum Wage	521
Example 14.6	The Decline of Private-Sector Unionism	527
Example 14.7	Wage Inequality—Have Computers Changed the Labor Market?	528
Example 15.1	The Value of Lost Earnings	537
Example 15.2	The Yields on Corporate Bonds	541
Example 15.3	Capital Investment in the Disposable Diaper Industry	548
Example 15.4	Choosing an Air Conditioner and a New Car	550
Example 15.5	How Depletable Are Depletable Resources?	554
Example 16.1	The Interdependence of International Markets	566
Example 16.2	The Effects of Automobile Import Quotas	588
Example 16.3	The Costs and Benefits of Special Protection	589
Example 17.1	Lemons in Major League Baseball	600
Example 17.2	Working into the Night	605

- Example 17.3 Reducing Moral Hazard—Warranties of Animal Health 608
- Example 17.4 Crisis in the Savings and Loan Industry 608
- Example 17.5 Managers of Nonprofit Hospitals as Agents 611
- Example 17.6 Efficiency Wages at Ford Motor Company 618
- Example 18.1 The Costs and Benefits of Reduced Sulfur
Dioxide Emissions 631
- Example 18.2 Emissions Trading and Clean Air 632
- Example 18.3 Regulating Municipal Solid Wastes 637
- Example 18.4 The Coase Theorem at Work 641
- Example 18.5 Crawfish Fishing in Louisiana 643
- Example 18.6 The Demand for Clear Air 647
- Example A.1 The Demand for Coal 661



PREFACE

For students who care about how the world works, microeconomics is one of the most relevant and interesting subjects they can study. A good grasp of microeconomics is vital for managerial decision making, for designing and understanding public policy, and more generally for appreciating how a modern economy functions.

We wrote this book, *Microeconomics*, because we believe that students need to be exposed to the new topics that have come to play a central role in microeconomics over the years—topics such as game theory and competitive strategy, the roles of uncertainty and information, and the analysis of pricing by firms with market power. We also felt that students need to be shown how microeconomics can help us to understand what goes on in the world and how it can be used as a practical tool for decision making. Microeconomics is an exciting and dynamic subject, but students need to be given an appreciation of its relevance and usefulness. They want and need a good understanding of how microeconomics can actually be used outside the classroom.

To respond to these needs, the fifth edition of *Microeconomics* provides a treatment of microeconomic theory that stresses its relevance and application to both managerial and public-policy decision making. This applied emphasis is accomplished by including 107 extended examples that cover such topics as the analysis of demand, cost, and market efficiency; the design of pricing strategies; investment and production decisions; and public policy analysis. Because of the importance that we attach to these examples, they are included in the flow of the text. (A complete list of the examples is included in the table of contents on pages ix–xxii.)

The coverage in the fifth edition of *Microeconomics* incorporates the dramatic changes that have occurred in the field in recent years. There has been growing interest in game theory and the strategic interactions of firms (Chapters 12 and 13), in the role and implications of uncertainty and asymmetric information (Chapters 5 and 17), in the pricing strategies of firms with market power (Chapters 10 and 11), and in the design of policies to deal efficiently with externalities such as environmental pollution (Chapter 18). These topics, which have only recently received attention in most books, are covered extensively here.

That the coverage in *Microeconomics* is comprehensive and up-to-date does not mean that it is “advanced” or difficult. We have worked hard to make the exposition clear and accessible as well as lively and engaging. We believe that the study of microeconomics should be enjoyable and stimulating. We hope that our book reflects this belief. Except for appendices and footnotes, *Microeconomics*

uses no calculus. As a result, it should be suitable for students with a broad range of backgrounds. (Those sections that are more demanding are marked with an asterisk and can be easily omitted.)

Changes in the Fifth Edition

Each new edition of this book has built on the success of prior editions by adding a number of new topics, by adding and updating examples, and by improving the exposition of existing materials. The fifth edition continues in that tradition. We have included a new section on auctions in Chapters 13 (Game Theory and Competitive Strategy), and we have expanded our coverage of supply-demand analysis in Chapter 2, as well as our coverage of cost in Chapters 7 and 8. In addition, we have added several new examples, and we have replaced a number of older examples with new ones.

In keeping with the preferences of many of our faithful users, we have not changed the chapter organization of the book. However, we have significantly revised portions of the first eight chapters, explaining some basic concepts in a more detailed and systematic way. Our primary goal in revising the book has been, as always, to make the text as clear, accessible, and engaging as possible.

The fifth edition of *Microeconomics*, like the fourth, is printed in four colors. As before, we have tried to use color to make the figures as clear and pedagogically effective as possible. In addition, we have added several new diagrams, and we have modified a number of existing diagrams to improve their accuracy and clarity.

This edition uses a larger text layout than earlier editions. This gave us the opportunity to add some new pedagogical devices. Key terms now appear in boldface and are defined in the margins of the text. Often, important ideas in microeconomics build on concepts that have been developed earlier in the text. In recognition of this fact, we have added a number of Concept Links in the margins, which explicitly direct the student to prior relevant materials.

Alternative Course Designs

The fifth edition of *Microeconomics* offers instructors substantial flexibility in course design. For a one-quarter or one-semester course stressing the basic core material, we would suggest using the following chapters and sections of chapters: 1, 2, 3, 4.1–4.4, 6, 7.1–7.4, 8, 9.1–9.3, 10, 11.1–11.3, 12, 14, 15.1–15.4, 18.1–18.2, and 18.5. A somewhat more ambitious course might also include parts of Chapters 5 and 16 and additional sections in Chapters 4, 7, and 9. To emphasize uncertainty and market failure, an instructor should also include substantial parts of Chapters 5 and 17.

Depending on one's interests and the goals of the course, other sections could be added or used to replace the materials listed above. A course emphasizing modern pricing theory and business strategy would include all of Chapters 11, 12, and 13 and the remaining sections of Chapter 15. A course in managerial economics might also include the appendixes to Chapters 4, 7, and 11 as well as the appendix on regression analysis at the end of the book. A course stressing welfare economics and public policy should include Chapter 16 and additional sections of Chapter 18.

Finally, we want to stress that those sections or subsections that are more demanding and/or peripheral to the core material have been marked with an asterisk. These sections can easily be omitted without detracting from the flow of the book.

Supplementary Materials

Ancillaries of an exceptionally high quality are available to instructors and students using the fifth edition of *Microeconomics*. The **Instructor's Manual**, prepared by Nora Underwood of the University of California, Davis, provides detailed solutions to all end-of-chapter Review Questions and Exercises. Each chapter also contains Teaching Tips to summarize key points and extra Review Questions with answers.

The **Test Bank**, prepared by John Crooker of Texas Tech University, contains over 2,000 multiple-choice and short-answer questions with solutions. It is designed for use with the Prentice Hall **Test Manager**, a computerized package that allows instructors to custom-design, save, and generate classroom tests.

A **PowerPoint Lecture Presentation**, created by Jeffrey Caldwell and Steven Smith, both of Rose State College, is available for the fifth edition and can be downloaded from the text Web site (www.prenhall.com/pindyck). Instructors can edit the detailed outlines and summaries to fit their own lecture presentations. A set of **Color Acetates** of the figures and selected tables from the text is available for instructors using the fifth edition of *Microeconomics*.

The **Study Guide**, prepared by Valerie Suslow of the University of Michigan and Jonathan Hamilton of the University of Florida, provides a wide variety of review materials and exercises for students. Each chapter contains a list of important concepts, chapter highlights, a concept review, problem sets, and a self-test quiz. Worked-out answers and solutions are provided for all exercises, problem sets, and self-test questions.

Prentice Hall's Learning on the Internet Partnership (myPHLIP)/Companion Web site (www.prenhall.com/pindyck) is a Web site with Internet exercises, activities, and resources related specifically to this text. New Internet resources are added every two weeks to provide both the student and the instructor with updated services. The site includes an **On-Line Study Guide**, prepared by Peter Zaleski of Villanova University, containing multiple-choice and essay questions. The On-Line Study Guide has a built-in grading feature that provides students with immediate feedback in the form of coaching comments.

For the instructor, the Web site offers such resources as the Syllabus Manager, answers to Current Events and Internet exercises, and a Faculty Lounge area with teaching archives and faculty chat rooms. From the Web site, instructors can also download supplements and lecture aids, including the Instructor's Manual and PowerPoint Lecture Presentation. Instructors should contact their Prentice Hall sales representative to get the necessary username and password to access the faculty resources on the site.

Prentice Hall provides faculty with Internet tools to help create on-line courses. It provides content and enhanced features to help instructors create full-length on-line courses or simply produce on-line supplementary materials to use in existing courses. Content is available on both WebCT and Blackboard platforms.

Acknowledgments

Because the fifth edition of *Microeconomics* has been the outgrowth of years of experience in the classroom, we owe a debt of gratitude to our students and to the colleagues with whom we often discuss microeconomics and its presentation. We have also had the help of capable research assistants. For the first four editions of the book, these included Walter Athier, Phillip Gibbs, Jamie Jue, Masaya Okoshi, Kathy O'Regan, Karen Randig, Subi Rangan, Deborah Senior, Ashesh Shah, and Wilson Tai. Kathy Hill helped with the art, while Assunta

Kent, Mary Knott, and Dawn Elliott Linahan provided secretarial assistance with the first edition. We especially want to thank Lynn Steele and Jay Tharp, who provided considerable editorial support for the second edition. Mark Glickman and Steve Wiggins assisted with the examples in the third edition, while Andrew Guest, Jeanette Sayre, and Lynn Steele provided valuable editorial support with the fourth edition.

Writing this book has been a painstaking and enjoyable process. At each stage we received exceptionally fine guidance from teachers of microeconomics throughout the country. After the first draft of the first edition of the book had been edited and reviewed, it was discussed at a two-day focus-group meeting in New York. This provided an opportunity to get ideas from instructors with a variety of backgrounds and perspectives. We would like to thank the following focus-group members for advice and criticism: Carl Davidson of Michigan State University; Richard Eastin of the University of Southern California; Judith Roberts of California State University, Long Beach; and Charles Strein of the University of Northern Iowa.

We would also like to thank all those who reviewed the first four editions at various stages of their evolution:

Jack Adams, University of Arkansas, Little Rock
 Sheri Aggarwal, Dartmouth College
 Ted Amato, University of North Carolina, Charlotte
 John J. Antel, University of Houston
 Kerry Back, Northwestern University
 Dale Ballou, University of Massachusetts, Amherst
 William Baxter, Stanford University
 James A. Brander, University of British Columbia
 Jeremy Bulow, Stanford University
 Winston Chang, State University of New York, Buffalo
 Henry Chappel, University of South Carolina
 Larry A. Chenault, Miami University
 Charles Clotfelter, Duke University
 Kathryn Combs, California State University, Los Angeles
 Richard Cornwall, Middlebury College
 John Coupe, University of Maine at Orono
 Jacques Cremer, Virginia Polytechnic Institute and State University
 Carl Davidson, Michigan State University
 Gilbert Davis, University of Michigan
 Arthur T. Denzau, Washington University
 Tran Dung, Wright State University
 Richard V. Eastin, University of Southern California
 Carl E. Enomoto, New Mexico State University
 Ray Farrow, Seattle University
 Gary Ferrier, Southern Methodist University
 Otis Gilley, Louisiana Tech University
 William H. Greene, New York University
 John Gross, University of Wisconsin at Milwaukee

Jonathan Hamilton, University of Florida
 Claire Hammond, Wake Forest University
 James Hartigan, University of Oklahoma
 George Heitman, Pennsylvania State University
 George E. Hoffer, Virginia Commonwealth University
 Robert Inman, The Wharton School, University of Pennsylvania
 Joyce Jacobsen, Rhodes College
 B. Patrick Joyce, Michigan Technological University
 David Kaserman, Auburn University
 Michael Kende, INSEAD, France
 Philip G. King, San Francisco State University
 Tetteh A. Kofi, University of San Francisco
 Anthony Krautman, DePaul University
 Leonard Lardaro, University of Rhode Island
 Peter Linneman, University of Pennsylvania
 R. Ashley Lyman, University of Idaho
 James MacDonald, Rensselaer Polytechnic Institute
 Wesley A. Magat, Duke University
 Anthony M. Marino, University of Southern Florida
 Richard D. McGrath, College of William and Mary
 David Mills, University of Virginia, Charlottesville
 Richard Mills, University of New Hampshire
 Jennifer Moll, Fairfield University
 Michael J. Moore, Duke University
 Julianne Nelson, Stern School of Business, New York University
 George Norman, Tufts University
 Daniel Orr, Virginia Polytechnic Institute and State University
 Sharon J. Pearson, University of Alberta, Edmonton
 Ivan P'ng, University of California, Los Angeles
 Michael Podgursky, University of Massachusetts, Amherst
 Charles Ratliff, Davidson College
 Judith Roberts, California State University, Long Beach
 Geoffrey Rothwell, Stanford University
 Nestor Ruiz, University of California, Davis
 Edward L. Sattler, Bradley University
 Roger Sherman, University of Virginia
 Nachum Sicherman, Columbia University
 Houston H. Stokes, University of Illinois, Chicago
 Richard W. Stratton, University of Akron
 Charles T. Strein, University of Northern Iowa
 Valerie Suslow, University of Michigan
 Abdul Turay, Radford University
 David Vrooman, St. Lawrence University
 Michael Wasylenko, Syracuse University
 Robert Whaples, Wake Forest University

Lawrence J. White, New York University
 Arthur Woolf, University of Vermont
 Chiou-nan Yeh, Alabama State University
 Joseph Ziegler, University of Arkansas, Fayetteville

We would like to thank the reviewers who provided comments and ideas that have contributed significantly to the Fifth Edition of *Microeconomics*:

Nii Adote Abrahams, Missouri Southern State College
 Victor Brajer, California State University, Fullerton
 Maxim Engers, University of Virginia
 Roger Frantz, San Diego State University
 Thomas A. Gresik, Pennsylvania State University
 Robert Lemke, Florida International University
 Lawrence Martin, Michigan State University
 John Makum Mbaku, Weber State University
 Charles Stuart, University of California, Santa Barbara
 Nora A. Underwood, University of California, Davis
 Peter Zaleski, Villanova University

Apart from the formal review process, we are especially grateful to Jean Andrews, Paul Anglin, J. C. K. Ash, Ernst Berndt, George Bittlingmayer, Severin Borenstein, Paul Carlin, Whewon Cho, Setio Angarro Dewo, Frank Fabozzi, Joseph Farrell, Frank Fisher, Jonathan Hamilton, Robert Inman, Joyce Jacobsen, Stacey Kole, Jeannette Mortensen, John Mullahy, Krishna Pendakur, Jeffrey Perloff, Ivan P'ng, A. Mitchell Polinsky, Judith Roberts, Geoffrey Rothwell, Garth Saloner, Joel Schrag, Daniel Siegel, Thomas Stoker, David Storey, and James Walker, who were kind enough to provide comments, criticisms, and suggestions as the various editions of this book developed.

Chapter 13 of the fifth edition contains new material on auctions, whose genesis owes much to the thoughtful comments and suggestions of Jonathan Hamilton, Preston McAfee, and Michael Williams. We especially want to thank Rashmi Khare and Nicola Stafford for their outstanding research assistance. Their help was critical in the development and updating of many of the examples and end-of-chapter exercises in this edition. We also want to thank Victor Brajer for carefully reviewing the page proofs for this edition. Finally, we are greatly indebted to Jeanette Sayre and Lynn Steele for their superb editorial work throughout the process of writing and producing the book.

We also wish to express our sincere thanks for the extraordinary effort those at Macmillan and Prentice Hall made in the development of the various editions of our book. Throughout the writing of the first edition, Bonnie Lieberman provided invaluable guidance and encouragement; Ken MacLeod kept the progress of the book on an even keel; Gerald Lombardi provided masterful editorial assistance and advice; and John Molyneux ably oversaw the book's production.

In the development of the second edition, we were fortunate to have the encouragement and support of David Boelio, and the organizational and editorial help of two Macmillan editors, Caroline Carney and Jill Lectka. The second edition also benefited greatly from the superb development editing of Gerald Lombardi, and from John Travis, who managed the book's production.

Jill Lectka and Denise Abbott were our editors for the third edition, and we benefited greatly from their input. We also want to thank Valerie Ashton, John Sollami, and Sharon Lee for their superb handling of the production of the third edition.

Leah Jewell was our editor for the fourth edition; her patience, thoughtfulness, and perseverance were greatly appreciated. We also want to thank our Production Editor, Dee Josephson, for managing the production process so effectively, and our Design Manager, Patricia Wosczyk, for her help with all aspects of the book's design.

Senior Economics Editor, Rod Banister, was our editor for the fifth edition. His focus, dedication, and thoughtfulness have been exemplary. We also appreciate the outstanding efforts of our Development Editor, Ron Librach; Managing Editor, Cynthia Regan; Design Manager, Pat Smythe; and Lorraine Castellano, who designed this edition. Likewise, we owe a debt of gratitude to many other professionals at Prentice Hall who played important roles in production and marketing. They include Editor in Chief, P. J. Boardman; Managing Editor, Gladys Soto; Assistant Editor, Holly Brown; Editorial Assistant, Marie McHale; Marketing Manager, Lori Braumburger; Senior Manufacturing Supervisor, Paul Smolenski; and Buyer, Lisa Babin.

R.S.P.

D.L.R.

PART 1

CHAPTERS

- 1 Preliminaries 3
- 2 The Basics of Supply and Demand 19

Introduction: Markets and Prices

PART 1 surveys the scope of microeconomics and introduces some basic concepts and tools. Chapter 1 discusses the range of problems that microeconomics addresses, and the kinds of answers it can provide. It also explains what a market is, how we determine the boundaries of a market, and how we measure market price.

Chapter 2 covers one of the most important tools of microeconomics: supply-demand analysis. We explain how a competitive market works and how supply and demand determine the prices and quantities of goods. We also show how supply-demand analysis can be used to determine the effects of changing market conditions, including government intervention.

CHAPTER 1

Preliminaries

Economics is divided into two main branches: microeconomics and macroeconomics. **Microeconomics** deals with the behavior of individual economic units. These units include consumers, workers, investors, owners of land, business firms—in fact, any individual or entity that plays a role in the functioning of our economy.¹ Microeconomics explains how and why these units make economic decisions. For example, it explains how consumers make purchasing decisions and how their choices are affected by changing prices and incomes. It also explains how firms decide how many workers to hire and how workers decide where to work and how much work to do.

Another important concern of microeconomics is how economic units interact to form larger units—markets and industries. Microeconomics helps us to understand, for example, why the American automobile industry developed the way it did and how producers and consumers interact in the market for automobiles. It explains how automobile prices are determined, how much automobile companies invest in new factories, and how many cars are produced each year. By studying the behavior and interaction of individual firms and consumers, microeconomics reveals how industries and markets operate and evolve, why they differ from one another, and how they are affected by government policies and global economic conditions.

By contrast, **macroeconomics** deals with aggregate economic quantities, such as the level and growth rate of national output, interest rates, unemployment, and inflation. But the boundary between macroeconomics and microeconomics has become less and less distinct in recent years. The reason is that macroeconomics also involves the analysis of markets—for example, the aggregate markets for goods and services, labor, and corporate bonds. To understand how these aggregate markets operate, we must first understand the behavior of the firms, consumers, workers, and investors who constitute them. Thus macroeconomists have become increasingly concerned with the microeconomic foundations of aggregate economic phenomena, and much of macroeconomics is actually an extension of microeconomic analysis.

¹ The prefix *micro-* is derived from the Greek word meaning “small.” However, many of the individual economic units that we will study are small only in relation to the U.S. economy as a whole. For example, the annual sales of General Motors, IBM, or Exxon are larger than the gross national products of many countries.

Chapter Outline

- 1.1 The Themes of Microeconomics 4
- 1.2 What Is a Market? 7
- 1.3 Real versus Nominal Prices 11
- 1.4 Why Study Microeconomics? 15

List of Examples

- 1.1 Markets for Prescription Drugs 10
- 1.2 The Price of Eggs and the Price of a College Education 12
- 1.3 The Minimum Wage 13

1.1 The Themes of Microeconomics

microeconomics Branch of economics that deals with the behavior of individual economic units—consumers, firms, workers, and investors—as well as the markets that these units comprise.

macroeconomics Branch of economics that deals with aggregate economic variables, such as the level and growth rate of national output, interest rates, unemployment, and inflation.

The Rolling Stones once said: “You can’t always get what you want.” This is true. For most people (even Mick Jagger), that there are limits to what you can have or do is a simple fact of life learned in early childhood. For economists, however, it can be an obsession.

Much of microeconomics is about limits—the limited incomes that consumers can spend on goods and services, the limited budgets and technical know-how that firms can use to produce things, and the limited number of hours in a week that workers can allocate to labor or leisure. But microeconomics is also about *ways to make the most of these limits*. More precisely, it is about *the allocation of scarce resources*. For example, microeconomics explains how consumers can best allocate their limited incomes to the various goods and services available for purchase. It explains how workers can best allocate their time to labor instead of leisure, or to one job instead of another. And it explains how firms can best allocate limited financial resources to hiring additional workers versus buying new machinery, and to producing one set of products versus another.

In a planned economy such as that of Cuba, North Korea, or the former Soviet Union, these allocation decisions are made mostly by the government. Firms are told what and how much to produce, and how to produce it; workers have little flexibility in choice of jobs, hours worked, or even where they live; and consumers typically have a very limited set of goods to choose from. As a result, many of the tools and concepts of microeconomics are of limited relevance in those countries.

In modern market economies, consumers, workers, and firms have much more flexibility and choice when it comes to allocating scarce resources. Microeconomics describes the *trade-offs* that consumers, workers, and firms face, and *shows how these trade-offs are best made*.

The idea of making optimal trade-offs is an important theme in microeconomics—one that you will encounter throughout this book. Let’s look at it in more detail.

Consumers Consumers have limited incomes, which can be spent on a wide variety of goods and services, or saved for the future. *Consumer theory*, the subject matter of Chapters 3, 4, and 5 of this book, describes how consumers, based on their preferences, maximize their well-being by trading off the purchase of more of some goods with the purchase of less of others. We will also see how consumers decide how much of their incomes to save, thereby trading off current consumption for future consumption.

Workers Workers also face constraints and make trade-offs. First, people must decide whether and when to enter the workforce. Because the kinds of jobs—and corresponding pay scales—available to a worker depend in part on educational attainment and accumulated skills, one must trade off working now (and earning an immediate income) with continued education (and the hope of earning a higher future income). Second, workers face trade-offs in their choice of employment. For example, while some people choose to work for large corporations that offer job security but limited potential for advancement, others prefer to work for small companies where there is more opportunity for advancement but less security. Finally, workers must sometimes decide how many hours per week they wish to work, thereby trading off labor for leisure.

Firms Firms also face limits in terms of the kinds of products that they can produce, and the resources available to produce them. The Ford Motor Company, for example, is very good at producing cars and trucks, but it does not have the ability to produce airplanes, computers, or pharmaceuticals. It is also constrained in terms of financial resources and the current production capacity of its factories. Given these constraints, Ford must decide how many of each type of vehicle to produce. If it wants to produce a larger total number of cars and trucks next year or the year after, it must decide whether to hire more workers, build new factories, or do both. The *theory of the firm*, the subject matter of Chapters 6 and 7, describes how these trade-offs can best be made.

A second important theme of microeconomics is the role of *prices*. All of the trade-offs described above are based on the prices faced by consumers, workers, or firms. For example, a consumer trades off beef for chicken based partly on his or her preferences for each one, but also on their prices. Likewise, workers trade off labor for leisure based in part on the “price” that they can get for their labor—i.e., the *wage*. And firms decide whether to hire more workers or purchase more machines based in part on wage rates and machine prices.

Microeconomics also describes how prices are determined. In a centrally planned economy, prices are set by the government. In a market economy, prices are determined by the interactions of consumers, workers, and firms. These interactions occur in *markets*—collections of buyers and sellers that together determine the price of a good. In the automobile market, for example, car prices are affected by competition among Ford, General Motors, Toyota, and other manufacturers, and also by the demands of consumers. The central role of markets is the third important theme of microeconomics. We will say more about the nature and operation of markets shortly.

Theories and Models

Like any science, economics is concerned with the *explanation* and *prediction* of observed phenomena. Why, for example, do firms tend to hire or lay off workers when the prices of their raw materials change? How many workers are likely to be hired or laid off by a firm or an industry if the price of raw materials increases by, say, 10 percent?

In economics, as in other sciences, explanation and prediction are based on *theories*. Theories are developed to explain observed phenomena in terms of a set of basic rules and assumptions. The *theory of the firm*, for example, begins with a simple assumption—firms try to maximize their profits. The theory uses this assumption to explain how firms choose the amounts of labor, capital, and raw materials that they use for production and the amount of output they produce. It also explains how these choices depend on the *prices* of inputs, such as labor, capital, and raw materials, and the prices that firms can receive for their outputs.

Economic theories are also the basis for making predictions. Thus the theory of the firm tells us whether a firm’s output level will increase or decrease in response to an increase in wage rates or a decrease in the price of raw materials. With the application of statistical and econometric techniques, theories can be used to construct models from which quantitative predictions can be made. A *model* is a mathematical representation, based on economic theory, of a firm, a market, or some other entity. For example, we might develop a model of a particular firm and use it to predict *by how much* the firm’s output level will change as a result of, say, a 10-percent drop in the price of raw materials.

Statistics and econometrics also let us measure the *accuracy* of our predictions. For example, suppose we predict that a 10-percent drop in the price of raw materials will lead to a 5-percent increase in output. Are we sure that the increase in output will be exactly 5 percent, or might it be somewhere between 3 and 7 percent? Quantifying the accuracy of a prediction can be as important as the prediction itself.

No theory, whether in economics, physics, or any other science, is perfectly correct. The usefulness and validity of a theory depend on whether it succeeds in explaining and predicting the set of phenomena that it is intended to explain and predict. Theories, therefore, are continually tested against observation. As a result of this testing, they are often modified or refined and occasionally even discarded. The process of testing and refining theories is central to the development of economics as a science.

When evaluating a theory, it is important to keep in mind that it is invariably imperfect. This is the case in every branch of science. In physics, for example, Boyle's law relates the volume, temperature, and pressure of a gas.² The law is based on the assumption that individual molecules of a gas behave as though they were tiny, elastic billiard balls. Physicists today know that gas molecules do not, in fact, always behave like billiard balls, which is why Boyle's law breaks down under extremes of pressure and temperature. Under most conditions, however, it does an excellent job of predicting how the temperature of a gas will change when the pressure and volume change, and it is therefore an essential tool for engineers and scientists.

The situation is much the same in economics. For example, firms do not maximize their profits all the time. Perhaps because of this, the theory of the firm has had only limited success in explaining certain aspects of firms' behavior, such as the timing of capital investment decisions. Nonetheless, the theory does explain a broad range of phenomena regarding the behavior, growth, and evolution of firms and industries, and so it has become an important tool for managers and policymakers.

Positive versus Normative Analysis

Microeconomics is concerned with both *positive* and *normative* questions. Positive questions deal with explanation and prediction, normative questions with what ought to be. Suppose the U.S. government imposes a quota on the import of foreign cars. What will happen to the price, production, and sales of cars? What impact will this policy change have on American consumers? On workers in the automobile industry? These questions belong to the realm of **positive analysis**: statements that describe relationships of cause and effect.

Positive analysis is central to microeconomics. As we explained above, theories are developed to explain phenomena, tested against observations, and used to construct models from which predictions are made. The use of economic theory for prediction is important both for the managers of firms and for public policy. Suppose the federal government is considering raising the tax on gasoline. The change would affect the price of gasoline, consumers' preferences for small or large cars, the amount of driving that people do, and so on. To plan sensibly,

² Robert Boyle (1627–1691) was a British chemist and physicist who discovered experimentally that pressure (P), volume (V), and temperature (T) were related in the following way: $PV = RT$, where R is a constant. Later, physicists derived this relationship as a consequence of the kinetic theory of gases, which describes the movement of gas molecules in statistical terms.

oil companies, automobile companies, producers of automobile parts, and firms in the tourist industry would all need to estimate the impact of the change. Government policymakers would also need quantitative estimates of the effects. They would want to determine the costs imposed on consumers (perhaps broken down by income categories); the effects on profits and employment in the oil, automobile, and tourist industries; and the amount of tax revenue likely to be collected each year.

Sometimes we want to go beyond explanation and prediction to ask such questions as "What is best?" This involves **normative analysis**, which is also important for both managers of firms and those making public policy. Again, consider a new tax on gasoline. Automobile companies would want to determine the best (profit-maximizing) mix of large and small cars to produce once the tax is in place. Specifically, how much money should be invested to make cars more fuel-efficient? For policymakers, the primary issue is likely to be whether the tax is in the public interest. The same policy objectives (say, an increase in tax revenues and a decrease in dependence on imported oil) might be met more cheaply with a different kind of tax, such as a tariff on imported oil.

Normative analysis is not only concerned with alternative policy options; it also involves the design of particular policy choices. For example, suppose it has been decided that a gasoline tax is desirable. Balancing costs and benefits, we then ask what is the optimal size of the tax.

Normative analysis is often supplemented by value judgments. For example, a comparison between a gasoline tax and an oil import tariff might conclude that the gasoline tax will be easier to administer but will have a greater impact on lower-income consumers. At that point, society must make a value judgment, weighing equity against economic efficiency.³ When value judgments are involved, microeconomics cannot tell us what the best policy is. However, it can clarify the trade-offs and thereby help to illuminate the issues and sharpen the debate.

1.2 What Is a Market?

We can divide individual economic units into two broad groups according to function—*buyers* and *sellers*. Buyers include consumers, who purchase goods and services; and firms, which buy labor, capital, and raw materials that they use to produce goods and services. Sellers include firms, which sell their goods and services; workers, who sell their labor services; and resource owners, who rent land or sell mineral resources to firms. Clearly, most people and most firms act as both buyers and sellers, but we will find it helpful to think of them as simply buyers when they are buying something, and sellers when they are selling something.

Together, buyers and sellers interact to form *markets*. A **market** is the collection of buyers and sellers that, through their actual or potential interactions, determine the price of a product or set of products. In the market for personal computers, for example, the buyers are business firms, households, and students; the sellers are

normative analysis Analysis examining questions of what ought to be.

market Collection of buyers and sellers that, through their actual or potential interactions, determine the price of a product or set of products.

positive analysis Analysis describing relationships of cause and effect.

³ Most of the value judgments involving economic policy boil down to just this trade-off—equity versus economic efficiency. This conflict and its implications are discussed clearly and in depth in Arthur M. Okun, *Equality and Efficiency: The Big Tradeoff* (Washington: Brookings Institution, 1975).

Compaq, IBM, Dell, Gateway, and a number of other firms. Note that a market includes more than an *industry*. An *industry* is a collection of firms that sell the same or closely related products. In effect, an industry is the supply side of the market.

market definition Determination of the buyers, sellers, and range of products that should be included in a particular market.

Economists are often concerned with **market definition**: which buyers and sellers should be included in a particular market. When defining a market, *potential* interactions of buyers and sellers can be just as important as *actual* ones. An example of this is the market for gold. A New Yorker who wants to buy gold is unlikely to travel to Zurich to do so. Most buyers of gold in New York will interact only with sellers in New York. But because the cost of transporting gold is small relative to its value, buyers of gold in New York *could* purchase their gold in Zurich if the prices there were significantly lower. Significant differences in the price of a commodity create a potential for **arbitrage**: buying at a low price in one location and selling at a higher price somewhere else. It is precisely this possibility of arbitrage which prevents the prices of gold in New York and Zurich from differing significantly and which creates a world market for gold.

arbitrage Practice of buying at a low price at one location and selling at a higher price in another.

Markets are at the center of economic activity, and many of the most interesting questions and issues in economics concern the functioning of markets. For example, why do only a few firms compete with one another in some markets, while in others a great many firms compete? Are consumers necessarily better off if there are many firms? If so, should the government intervene in markets with only a few firms? Why have prices in some markets risen or fallen rapidly, while in other markets prices have hardly changed at all? And which markets offer the best opportunities for an entrepreneur thinking of going into business?

Competitive versus Noncompetitive Markets

In this book, we study the behavior of both competitive and noncompetitive markets. A *perfectly competitive market* has many buyers and sellers, so that no single buyer or seller has a significant impact on price. Most agricultural markets are close to being perfectly competitive. For example, thousands of farmers produce wheat, which thousands of buyers purchase to produce flour and other products. As a result, no single farmer and no single buyer can significantly affect the price of wheat.

Many other markets are competitive enough to be treated as if they were perfectly competitive. The world market for copper, for example, contains a few dozen major producers. That number is enough for the impact on price to be negligible if any one producer goes out of business. The same is true for many other natural resource markets, such as those for coal, iron, tin, or lumber.

Other markets containing a small number of producers may still be treated as competitive for purposes of analysis. For example, the U.S. airline industry contains several dozen firms, but most routes are served by only a few firms. Nonetheless, because competition among those firms is often fierce, for some purposes the market can be treated as competitive. Finally, some markets contain many producers but are *noncompetitive*; that is, individual firms can jointly affect the price. The world oil market is one example. Since the early 1970s, that market has been dominated by the OPEC cartel. (A *cartel* is a group of producers that acts collectively.)

Market Price

Markets make possible transactions between buyers and sellers. Quantities of a good are sold at specific prices. In a perfectly competitive market, a single price—the **market price**—will usually prevail. The price of wheat in Kansas

market price Price prevailing in a competitive market.

City and the price of gold in New York are two examples. These prices are usually easy to measure. For example, you can find the price of corn, wheat, or gold each day in the business section of a newspaper.

In markets that are not perfectly competitive, different firms might charge different prices for the same product. This might happen because one firm is trying to win customers from its competitors, or because customers have brand loyalties that allow some firms to charge higher prices than others. For example, two brands of laundry detergent might be sold in the same supermarket at different prices. Or two supermarkets in the same town might sell the same brand of laundry detergent at different prices. In cases such as this, when we refer to the market price, we will mean the price averaged across brands or supermarkets.

The market prices of most goods will fluctuate over time, and for many goods the fluctuations can be rapid. This is particularly true for goods sold in competitive markets. The stock market, for example, is highly competitive because there are typically many buyers and sellers for any one stock. As anyone who has invested in the stock market knows, the price of any particular stock fluctuates from minute to minute and can rise or fall substantially during a single day. Likewise, the prices of commodities such as wheat, soybeans, coffee, oil, gold, silver, and lumber can rise or fall dramatically in a day or a week.

Market Definition—The Extent of a Market

As we saw, *market definition* identifies which buyers and sellers should be included in a given market. However, to determine which buyers and sellers to include, we must first determine the *extent of the market*. The **extent of a market** refers to its *boundaries*, both *geographically* and in terms of the *range of products* to be included in it.

extent of a market Boundaries of a market, both geographical and in terms of range of products produced and sold within it.

When we refer to the market for gasoline, for example, we must be clear about its geographic boundaries. Are we referring to downtown Los Angeles, southern California, or the entire United States? We must also be clear about the range of products to which we are referring. Should regular-octane and high-octane premium gasoline be included in the same market? Leaded and unleaded gasoline? Gasoline and diesel fuel?

For some goods, it makes sense to talk about a market only in terms of very restrictive geographic boundaries. Housing is a good example. Most people who work in downtown Chicago will look for housing within commuting distance. They will not look at homes 200 or 300 miles away, even though those homes might be much cheaper. And homes (together with the land they are sitting on) 200 miles away cannot be easily moved closer to Chicago. Thus the housing market in Chicago is separate and distinct from, say, those in Cleveland, Houston, Atlanta, or Philadelphia. Likewise, retail gasoline markets, though less limited geographically, are still regional because of the expense of shipping gasoline over long distances. Thus the market for gasoline in southern California is distinct from that in northern Illinois. On the other hand, as we mentioned earlier, gold is bought and sold in a world market; the possibility of arbitrage prevents the price from differing significantly from one location to another.

We must also think carefully about the range of products to include in a market. For example, there is a market for 35-millimeter single-lens reflex (SLR) cameras, and many brands compete in that market. But what about Polaroid instant cameras? Should they be considered part of the same market? Probably not, because they are used for different purposes and so do not compete with

SLR cameras. Gasoline is another example. Regular and premium octane gasolines might be considered part of the same market because most consumers can use either. Diesel fuel, however, is not part of this market because cars that use regular gasoline cannot use diesel fuel, and vice versa.⁴

Market definition is important for a number of reasons. A company, for example, must understand who its actual and potential competitors are for the various products it now sells or might sell in the future. It must also know the product-characteristic boundaries and geographical boundaries of its market in order to be able to set price, determine advertising budgets, and make capital investment decisions. Market definition is likewise important for public-policy decisions. Should the government allow a merger or acquisition involving companies that produce similar products, or should it challenge it? The answer depends on the impact of that merger or acquisition on future competition and prices, and often this can be evaluated only by defining the market.

EXAMPLE 1.1 Markets for Prescription Drugs

The development of a new drug by a pharmaceutical company is an expensive venture. It begins with large expenditures on research and development, then requires various stages of laboratory and clinical testing, and, if the new drug is finally approved, marketing, production, and sales. At that point, the firm faces the important problem of determining the price of the new drug. Pricing depends on the preferences and medical needs of the consumers who will be buying the drug, the characteristics of the drug, and the number and characteristics of *competing* drugs. Pricing a new drug, therefore, requires a good understanding of the market in which it will be sold.

In the pharmaceutical industry, market boundaries are sometimes easy to determine, and sometimes not so easy to determine. Markets are usually defined in terms of *therapeutic classes* of drugs. For example, there is a market for *antiulcer drugs* that is very clearly defined. Until a few years ago, there were four competitors in the market: Tagamet (produced by Smithkline-Beecham), Zantac (produced by Glaxo), Axid (produced by Eli Lilly), and Pepcid (produced by Merck). All four drugs work in roughly the same way: They cause the stomach to produce less hydrochloric acid. They differ slightly in terms of their side effects and their interactions with other drugs that a patient might be taking, but in most cases they could be readily substituted for each other.⁵

Another example of a clearly defined pharmaceutical market is the market for *anticholesterol* drugs. There are four major products in the market: Merck's Mevacor has about 50 percent of the market. Pravachol (Bristol-Myers-Squibb) and Zocor (also Merck) each have about 20 percent, and Lescol (Sandoz) about 10 percent. These drugs all do pretty much the same thing (reduce blood cholesterol levels) and work in pretty much the same way. While their side effects

⁴ How can we determine the extent of a market? Since the market is where the price of a good is established, one approach focuses on market prices. We ask whether product prices in different geographic regions (or for different product types) are approximately the same, or whether they tend to move together. If either is the case, we place them in the same market. For a more detailed discussion, see George J. Stigler and Robert A. Sherwin, "The Extent of the Market," *Journal of Law and Economics* 27 (October 1985): 555–85.

⁵ As we will discuss in Example 10.1, more recently Prilosec entered the market, and by 1997 became the largest selling drug in the world. It is also an antiulcer drug, but works on a different biochemical mechanism.

and interactions differ somewhat, they are all close substitutes. Thus when Merck sets the price of Mevacor, it must be concerned not only with the willingness of patients (and their insurance companies) to pay, but also with the prices and characteristics of the three competing drugs. Likewise, a drug company that is considering whether to develop a new anticholesterol drug knows that if it commits itself to the investment and succeeds, it will have to compete with the four existing drugs. The company can use this information to project its potential revenues from the new drug, and thereby evaluate the investment.

Sometimes pharmaceutical market boundaries are more ambiguous. Consider painkillers, a category that includes aspirin, acetaminophen (sold under the brand name Tylenol but also sold generically), ibuprofen (sold under such brand names as Motrin and Advil, but also sold generically), naproxen (sold by prescription, but also sold over the counter by the brand name Aleve), and Voltaren (a more powerful prescription drug produced by Novartis). There are many types of painkillers, and some work better than others for certain types of pain (e.g., headaches, arthritis, muscle aches, etc.). Side effects likewise differ. While some types of painkillers are used more frequently for certain symptoms or conditions, there is considerable spillover. For example, depending on the severity of the pain and the pain tolerance of the patient, a toothache might be treated with any of the painkillers listed above. This substitutability makes the boundaries of the painkiller market difficult to define.

1.3 Real versus Nominal Prices

We often want to compare the price of a good today with what it was in the past or is likely to be in the future. To make such a comparison meaningful, we need to measure prices relative to the *overall price level*. In absolute terms, the price of a dozen eggs is many times higher today than it was 50 years ago. Relative to prices overall, however, it is actually lower. Therefore, we must be careful to correct for inflation when comparing prices across time. This means measuring prices in *real* rather than *nominal* terms.

The **nominal price** of a good (sometimes called its "current-dollar" price) is just its absolute price. For example, the nominal price of a quart of milk was about 40 cents in 1970, about 65 cents in 1980, and about \$1.05 in 1999. These are the prices you would have seen in supermarkets in those years. The **real price** of a good (sometimes called its "constant-dollar" price) is the price relative to an aggregate measure of prices. In other words, it is the price adjusted for inflation.

The aggregate measure most often used is the **Consumer Price Index (CPI)**. The CPI is calculated by the U.S. Bureau of Labor Statistics and is published monthly. It records how the cost of a large market basket of goods purchased by a "typical" consumer in some base year changes over time. (Currently the base year is 1983.) Percentage changes in the CPI measure the rate of inflation in the economy.⁶

⁶ Because the market basket is fixed, the CPI can tend to overstate inflation. The reason is that when the prices of some goods rise substantially, consumers will shift some of their purchases to goods whose prices have not risen as much, and the CPI ignores this phenomenon. We will discuss this in Chapter 3.

nominal price Absolute price of a good, unadjusted for inflation.

real price Price of a good relative to an aggregate measure of prices; price adjusted for inflation.

Consumer Price Index Measure of the aggregate price level.

After correcting for inflation, do we find that milk was more expensive in 1999 than in 1970? To find out, let's calculate the 1999 price of milk in terms of 1970 dollars. The CPI was 38.8 in 1970 and rose to about 167 in 1999.⁷ (There was considerable inflation in the United States during the 1970s and early 1980s.) In 1970 dollars, the price of milk was

$$\frac{38.8}{167} \times \$1.05 = \$0.24$$

In real terms, therefore, the price of milk was lower in 1999 than it was in 1970. Put another way, the nominal price of milk went up by about 162 percent, but the CPI went up 330 percent. Relative to the aggregate price level, milk prices fell.

In this book, we will usually be concerned with real rather than nominal prices because consumer choices involve analyses of price comparisons. These relative prices can most easily be evaluated if there is a common basis of comparison. Stating all prices in real terms achieves this objective. Thus, even though we will often measure prices in dollars, we will be thinking in terms of the real purchasing power of those dollars.

EXAMPLE 1.2 The Price of Eggs and the Price of a College Education

In 1970, Grade A large eggs cost about 61 cents a dozen. In the same year, the average annual cost of a college education at a private four-year college, including room and board, was about \$2,530. By 1998, the price of eggs had risen to \$1.04 a dozen, and the average cost of a college education was \$19,213. In real terms, were eggs more expensive in 1998 than in 1970? Had a college education become more expensive?

Table 1.1 shows the nominal price of eggs, the nominal cost of a college education, and the CPI for 1970–1998. (The CPI is based on 1983 = 100.) Also

	1970	1975	1980	1985	1990	1998
Consumer Price Index	38.8	53.8	82.4	107.6	130.7	163.0
Nominal Prices						
Grade A large eggs	\$0.61	\$0.77	\$0.84	\$0.80	\$1.01	\$1.04
College education	2530	3403	4912	8156	12,800	19,213
Real Prices (\$1970)						
Grade A large eggs	\$0.61	\$0.56	\$0.40	\$0.29	\$0.30	\$0.25
College education	2530	2454	2313	2941	3800	4573

⁷ Two good sources of data on the national economy are the *Economic Report of the President* and the *Statistical Abstract of the United States*. Both are published annually and are available from the U.S. Government Printing Office.

shown are the *real* prices of eggs and a college education in 1970 dollars, calculated as follows:

$$\text{Real price of eggs in 1975} = \frac{\text{CPI}_{1970}}{\text{CPI}_{1975}} \times \text{nominal price in 1975}$$

$$\text{Real price of eggs in 1980} = \frac{\text{CPI}_{1970}}{\text{CPI}_{1980}} \times \text{nominal price in 1980}$$

and so forth.

The table shows clearly that the real cost of a college education rose (by 81 percent) during this period, while the real cost of eggs fell (by 59 percent). It is these relative changes in prices that are important for the choices that consumers must make, not the fact that both eggs and college cost more in dollars today than they did in 1970.

In the table, we calculated real prices in terms of 1970 dollars, but we could have just as easily calculated them in terms of dollars of some other base year. For example, suppose we want to calculate the real price of eggs in 1980 dollars. Then:

$$\text{Real price of eggs in 1975} = \frac{\text{CPI}_{1980}}{\text{CPI}_{1975}} \times \text{nominal price in 1975}$$

$$\text{Real price of eggs in 1985} = \frac{\text{CPI}_{1980}}{\text{CPI}_{1985}} \times \text{nominal price in 1985}$$

and so forth. By going through the calculations, you can check to see that in terms of 1980 dollars, the real price of eggs was \$1.30 in 1970, \$1.18 in 1975, 84 cents in 1980, 61 cents in 1985, 64 cents in 1990, and 53 cents in 1998. You will also see that the percentage declines in real price are the same no matter which base year we use.⁸

EXAMPLE 1.3 The Minimum Wage

The federal minimum wage—first instituted in 1938 at a level of 25 cents per hour—has been increased periodically over the years. From 1981 through 1989, for example, it was \$3.35 an hour and was raised to \$4.25 an hour in 1990. In 1996, after much deliberation and debate, Congress voted to raise the minimum wage to \$4.70 in 1996 and then to \$5.15 in 1997.⁹

Figure 1.1 shows the minimum wage from 1938 through 1999, both in nominal terms and in 1996 constant dollars. Note that although the legislated minimum wage has steadily increased, in real terms the minimum wage today is not very different from what it was in the 1950s.

Nonetheless, the 1996 decision to increase the minimum wage was a difficult one. Although the higher minimum wage would provide a better standard of

⁸ You can get World Wide Web data on the cost of a college education at <http://www.collegeboard.org> and on the price of eggs at <http://www.econ.ag.gov/briefing/foodmark/retail/data/meat/eggs/gra.htm>

⁹ Some states also have minimum wages that are higher than the federal minimum wage. You can learn more about the minimum wage at this Web site: http://www.dol.gov/dol/esa/public/minwage_main.htm

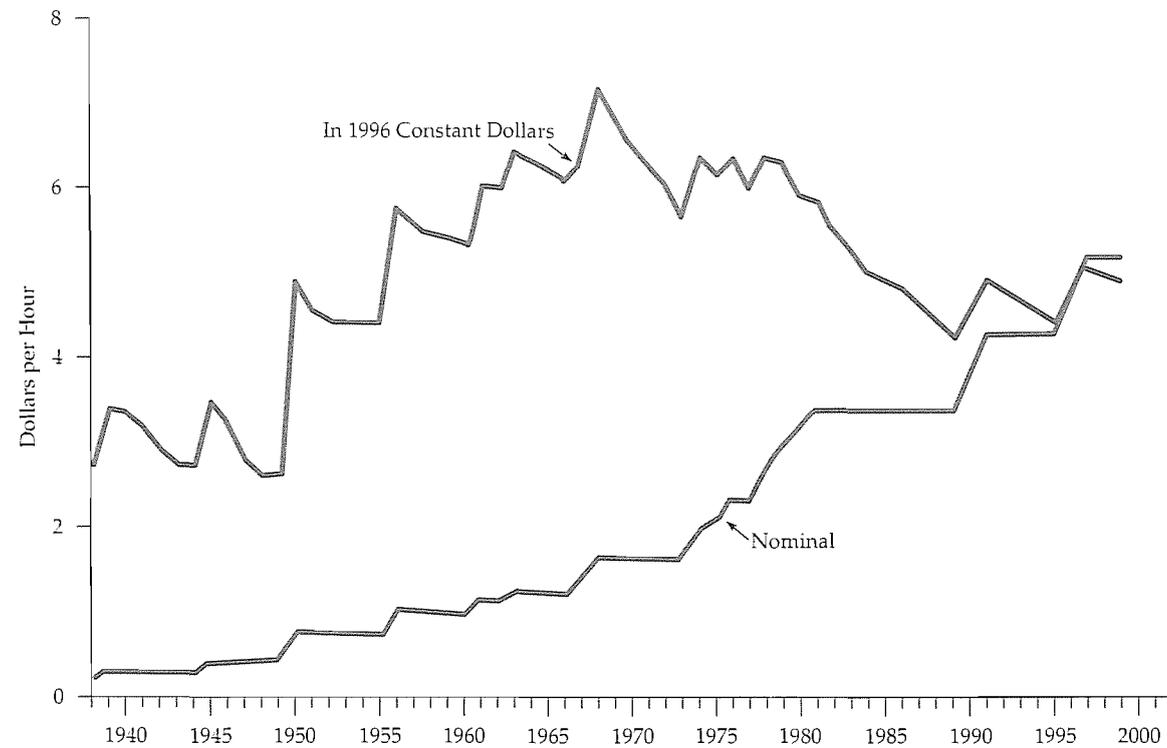


FIGURE 1.1 The Minimum Wage

In nominal terms, the minimum wage has increased steadily over the past 60 years. However, in real terms its 1999 level is below that of the 1970s.

living for those workers who had been paid below the minimum, some analysts feared that it would also lead to increased unemployment among young and unskilled workers. The decision to increase the minimum wage, therefore, raises both normative and positive issues. The normative issue is whether any loss of teenage and low-skilled jobs is outweighed by two factors: (1) the direct benefits to those workers who now earn more as a result; and (2) any indirect benefits to other workers whose wages might be increased along with the wages of those at the bottom of the pay scale.

An important positive issue is how many fewer workers (if any) would be able to get jobs with a higher minimum wage. As we will see in Chapter 14, this issue is still hotly debated. Statistical studies have suggested that an increase in the minimum wage of about 10 percent would increase teenage unemployment by 1 to 2 percent. (The actual increase from \$4.25 to \$5.15 represents a \$0.90/\$4.25 or 21 percent increase.) However, one recent review of the evidence questions whether there are any significant unemployment effects at all.¹⁰

¹⁰ The first study is David Neumark and William Wascher, "Employment Effects of Minimum and Subminimum Wages: Panel Data on State Minimum Wage Laws," *Industrial and Labor Relations Review* 46 (October 1992): 55–81. A review of the literature appears in David Card and Alan Krueger, *Myth and Measurement: The New Economics of the Minimum Wage* (Princeton: Princeton University Press, 1995).

1.4 Why Study Microeconomics?

We think that after reading this book you will have no doubt about the importance and broad applicability of microeconomics. In fact, one of our major goals is to show you how to apply microeconomic principles to actual decision-making problems. Nonetheless, some extra motivation early on never hurts. Here are two examples that not only show the use of microeconomics in practice but also provide a preview of this book.

Corporate Decision Making: Ford's Sport Utility Vehicles

By the mid 1990s, the Ford Explorer had become the best-selling sport utility vehicle (SUV) in the United States. Then, in 1997, Ford introduced the Expedition—a newly designed, larger, and roomier SUV. This car was also a huge success, and contributed significantly to Ford's profits. The success of these cars led Ford to introduce an even larger and heavier SUV in 1999—the Excursion. The design and efficient production of these cars involved not only some impressive engineering advances, but a lot of economics as well.

First, Ford had to think carefully about how the public would react to the design and performance of its new products. How strong would demand be initially, and how fast would it grow? How would demand depend on the prices that Ford charged? Understanding consumer preferences and trade-offs and predicting demand and its responsiveness to price are essential to Ford and every other automobile manufacturer. (We discuss consumer preferences and demand in Chapters 3, 4, and 5.)

Next, Ford had to be concerned with the cost of manufacturing these cars. How high would production costs be? How would costs depend on the number of cars produced each year? How would union wage negotiations or the prices of steel and other raw materials affect costs? How much and how fast would costs decline as managers and workers gained experience with the production process? And to maximize profits, how many of these cars should Ford plan to produce each year? (We discuss production and cost in Chapters 6 and 7 and the profit-maximizing choice of output in Chapter 8.)

Ford also had to design a pricing strategy and consider how competitors would react to it. For example, should Ford charge a low price for the basic stripped-down version of the Explorer but high prices for individual options, such as leather seats? Or would it be more profitable to make these options "standard" items and charge a higher price for the whole package? Whatever strategy Ford chose, how were competitors likely to react? Would DaimlerChrysler try to undercut Ford by lowering the price of its Jeep Grand Cherokee? Might Ford be able to deter DaimlerChrysler or GM from lowering prices by threatening to respond with its own price cuts? (We discuss pricing in Chapters 10 and 11 and competitive strategy in Chapters 12 and 13.)

Because its SUV product line required large investments in new capital equipment, Ford had to consider both the risks and possible outcomes of its decisions. Some of this risk was due to uncertainty over the future price of gasoline (higher gasoline prices would reduce the demand for heavy vehicles). Some was due to uncertainty over the wages that Ford would have to pay its workers. What would happen if world oil prices doubled or tripled, or if the U.S. government imposed a heavy tax on gasoline? How much bargaining power would the

unions have, and how might union demands affect wage rates? How should Ford take these uncertainties into account when making investment decisions? (Commodity markets and the effects of taxes are discussed in Chapters 2 and 9. Labor markets and union power are discussed in Chapter 14. Investment decisions and the role of uncertainty are discussed in Chapters 5 and 15.)

Ford also had to worry about organizational problems. Ford is an integrated firm in which separate divisions produce engines and parts and then assemble finished cars. How should managers of different divisions be rewarded? What price should the assembly division be charged for engines that it receives from another division? Should all parts be obtained from the upstream divisions, or should some be purchased from outside firms? (We discuss internal pricing and organizational incentives for the integrated firm in Chapters 11 and 17.)

Finally, Ford had to think about its relationship to the government and the effects of regulatory policies. For example, all of Ford's cars must meet federal emissions standards, and production-line operations must comply with health and safety regulations. How might these regulations and standards change over time? How might they affect costs and profits? (We discuss the role of government in limiting pollution and promoting health and safety in Chapter 18.)

Public Policy Design: Automobile Emission Standards for the Twenty-first Century

In 1970, the Federal Clean Air Act imposed strict tailpipe emission standards on new automobiles. These standards have become increasingly stringent—the 1970 levels of nitrogen oxides, hydrocarbons, and carbon monoxide emitted by automobiles had been reduced by about 90 percent by 1999. Now, as the number of cars on the roads keeps increasing, the government must consider how stringent these standards should be in the coming years.

The design of a program like the Clean Air Act involves a careful analysis of the ecological and health effects of auto emissions. But it also involves a good deal of economics. First, the government must evaluate the monetary impact of the program on consumers. Emission standards affect the cost both of purchasing a car (catalytic converters would be necessary, which would raise the cost of cars) and of operating it (gas mileage would be lower, and converters would have to be repaired and maintained). Because consumers ultimately bear much of this added cost, it is important to know how it affects their standards of living. This means analyzing consumer preferences and demand. For example, would consumers drive less and spend more of their income on other goods? If so, would they be nearly as well off? (Consumer preferences and demand are discussed in Chapters 3 and 4.)

To answer these questions, the government must determine how new standards will affect the cost of producing cars. Might automobile producers minimize cost increases by using new lightweight materials? (Production and cost are discussed in Chapters 6 and 7.) Then the government needs to know how changes in production costs will affect the production levels and prices of new automobiles. Are the additional costs absorbed or passed on to consumers in the form of higher prices? (Output determination is discussed in Chapter 8, and pricing in Chapters 10 through 13.)

Finally, the government must ask why the problems related to air pollution are not solved by our market-oriented economy. The answer is that much of the cost of air pollution is external to the firm. If firms do not find it in their self-interest to deal adequately with auto emissions, what is the best way to alter their incentives? Should standards be set, or is it more economical to impose air

pollution fees? How do we decide what people will pay to clean up the environment when there is no explicit market for clean air? Is the political process likely to solve these problems? The ultimate question is whether the auto emissions control program makes sense on a cost-benefit basis. Are the aesthetic, health, and other benefits of clean air worth the higher cost of automobiles? (These problems are discussed in Chapter 18.)

These are just two examples of how microeconomics can be applied in the arenas of private and public policy decision making. You will see many more applications as you read this book.

SUMMARY

1. Microeconomics is concerned with the decisions made by small economic units—consumers, workers, investors, owners of resources, and business firms. It is also concerned with the interaction of consumers and firms to form markets and industries.
2. Microeconomics relies heavily on the use of theory, which can (by simplification) help to explain how economic units behave and predict what behavior will occur in the future. Models are mathematical representations of theories that can help in this explanation and prediction process.
3. Microeconomics is concerned with positive questions that have to do with the explanation and prediction of phenomena. But microeconomics is also important for normative analysis, in which we ask what choices are best—for a firm or for society as a whole. Normative analyses must often be combined with individual value judgments because issues of equity and fairness as well as of economic efficiency may be involved.
4. A *market* refers to a collection of buyers and sellers who interact, and to the possibility for sales and purchases that results from that interaction. Microeconomics involves the study of both perfectly competitive markets, in which no single buyer or seller has an impact on price, and noncompetitive markets, in which individual entities can affect price.
5. The market price is established by the interaction of buyers and sellers. In a perfectly competitive market, a single price will usually prevail. In markets that are not perfectly competitive, different sellers might charge different prices. In this case, the market price refers to the average prevailing price.
6. When discussing a market, we must be clear about its extent in terms of both its geographic boundaries and the range of products to be included in it. Some markets (e.g., housing) are highly localized, whereas others (e.g., gold) are global in nature.
7. To eliminate the effects of inflation, we measure real (or constant-dollar) prices, rather than nominal (or current-dollar) prices. Real prices use an aggregate price index, such as the CPI, to correct for inflation.

QUESTIONS FOR REVIEW

1. It is often said that a good theory is one that can in principle be refuted by an empirical, data-oriented study. Explain why a theory that cannot be evaluated empirically is not a good theory.
2. Which of the following two statements involves positive economic analysis and which normative? How do the two kinds of analysis differ?
 - a. Gasoline rationing (allocating to each individual a maximum amount of gasoline that can be purchased each year) is a poor social policy because it interferes with the workings of the competitive market system.
 - b. Gasoline rationing is a policy under which more people are made worse off than are made better off.
3. Suppose the price of unleaded regular octane gasoline were 20 cents per gallon higher in New Jersey than in Oklahoma. Do you think there would be an opportunity for arbitrage (i.e., that firms could buy gas in Oklahoma and then sell it at a profit in Jersey)? Why or why not?
4. In Example 1.2, what economic forces explain why the real price of eggs has fallen, while the real price of a college education has increased? How have these changes affected consumer choices?

5. Suppose that the Japanese yen rises against the U.S. dollar; that is, it will take more dollars to buy any given amount of Japanese yen. Explain why this increase simultaneously increases the real price of Japanese cars for U.S. consumers and lowers the real price of U.S. automobiles for Japanese consumers.
6. The price of long-distance telephone service fell from 40 cents per minute in 1996 to 22 cents per minute in 1999, a 45-percent (18 cents/40 cents) decrease. The Consumer Price Index increased by 10 percent over this period. What happened to the real price of telephone service?

EXERCISES

1. Decide whether each of the following statements is true or false and explain why:
- Fast-food chains like McDonald's, Burger King, and Wendy's operate all over the United States. Therefore the market for fast food is a national market.
 - People generally buy clothing in the city in which they live. Therefore there is a clothing market in, say, Atlanta that is distinct from the clothing market in Los Angeles.
 - Some consumers strongly prefer Pepsi and some strongly prefer Coke. Therefore there is no single market for colas.
2. The following table shows the average retail price of milk and the Consumer Price Index from 1980 to 1998.
- | | 1980 | 1985 | 1990 | 1995 | 1998 |
|---|--------|--------|--------|--------|--------|
| CPI | 100 | 130.58 | 158.62 | 184.95 | 197.82 |
| Retail price of milk (fresh, whole, 1/2 gal.) | \$1.05 | \$1.13 | \$1.39 | \$1.48 | \$1.61 |
- Calculate the real price of milk in 1980 dollars. Has the real price increased/decreased/stayed the same since 1980?
 - What is the percentage change in the real price (1980 dollars) from 1980 to 1998?
 - Convert the CPI into 1990 = 100 and determine the real price of milk in 1990 dollars.
 - What is the percentage change in real price (1990 dollars) from 1980 to 1998? Compare this with your answer in (b). What do you notice? Explain.
3. At the time this book went to print, the minimum wage was \$5.15. To find the current value of the CPI, go to <http://www.bls.gov/top20.html>. Click on Consumer Price Index—All Urban Consumers (Current Series) and select U.S. All items. This will give you the CPI from 1913 to the present.
- With these values, calculate the current real minimum wage in 1990 dollars.
 - What is the percentage change in the real minimum wage from 1985 to the present, stated in real 1990 dollars?

CHAPTER 2

Chapter Outline

- 2.1 Supply and Demand 20
- 2.2 The Market Mechanism 23
- 2.3 Changes in Market Equilibrium 24
- 2.4 Elasticities of Supply and Demand 30
- 2.5 Short-Run versus Long-Run Elasticities 35
- *2.6 Understanding and Predicting the Effects of Changing Market Conditions 44
- 2.7 Effects of Government Intervention—Price Controls 53

List of Examples

- 2.1 The Price of Eggs and the Price of a College Education Revisited 26
- 2.2 Wage Inequality in the United States 27
- 2.3 The Long-Run Behavior of Natural Resource Prices 28
- 2.4 The Market for Wheat 33
- 2.5 The Demand for Gasoline and Automobiles 39
- 2.6 The Weather in Brazil and the Price of Coffee in New York 41
- 2.7 Declining Demand and the Behavior of Copper Prices 47
- 2.8 Upheaval in the World Oil Market 49
- 2.9 Price Controls and Natural Gas Shortages 54

The Basics of Supply and Demand

One of the best ways to appreciate the relevance of economics is to begin with the basics of supply and demand. Supply-demand analysis is a fundamental and powerful tool that can be applied to a wide variety of interesting and important problems. To name a few:

- Understanding and predicting how changing world economic conditions affect market price and production
- Evaluating the impact of government price controls, minimum wages, price supports, and production incentives
- Determining how taxes, subsidies, tariffs, and import quotas affect consumers and producers

We begin with a review of how supply and demand curves are used to describe the *market mechanism*. Without government intervention (e.g., through the imposition of price controls or some other regulatory policy), supply and demand will come into equilibrium to determine both the market price of a good and the total quantity produced. What that price and quantity will be depends on the particular characteristics of supply and demand. Variations of price and quantity over time depend on the ways in which supply and demand respond to other economic variables, such as aggregate economic activity and labor costs, which are themselves changing.

We will, therefore, discuss the characteristics of supply and demand and show how those characteristics may differ from one market to another. Then we can begin to use supply and demand curves to understand a variety of phenomena—for example, why the prices of some basic commodities have fallen steadily over a long period while the prices of others have experienced sharp gyrations; why shortages occur in certain markets; and why announcements about plans for future government policies or predictions about future economic conditions can affect markets well before those policies or conditions become reality.

Besides understanding *qualitatively* how market price and quantity are determined and how they can vary over time, it is also important to learn how they can be analyzed *quantitatively*. We will see how simple “back of the envelope” calculations can be used to analyze and predict evolving market conditions.

We will also show how markets respond both to domestic and international macroeconomic fluctuations and to the effects of government interventions. We will try to convey this understanding through simple examples and by urging you to work through some exercises at the end of the chapter.

2.1 Supply and Demand

The basic model of supply and demand is the workhorse of microeconomics. It helps us understand why and how prices change, and what happens when the government intervenes in a market. The supply-demand model combines two important concepts: a *supply curve* and a *demand curve*. It is important to understand precisely what these curves represent.

The Supply Curve

supply curve Relationship between the quantity of a good that producers are willing to sell and the price of the good.

The **supply curve** shows the quantity of a good that producers are willing to sell at a given price, holding constant any other factors that might affect the quantity supplied. The curve labeled S in Figure 2.1 illustrates this. The vertical axis of the graph shows the price of a good, P , measured in dollars per unit. This is the price that sellers receive for a given quantity supplied. The horizontal axis shows the total quantity supplied, Q , measured in the number of units per period.

The supply curve is thus a relationship between the quantity supplied and the price. We can write this relationship as an equation:

$$Q_S = Q_S(P)$$

or we can draw it graphically, as we have done in Figure 2.1.

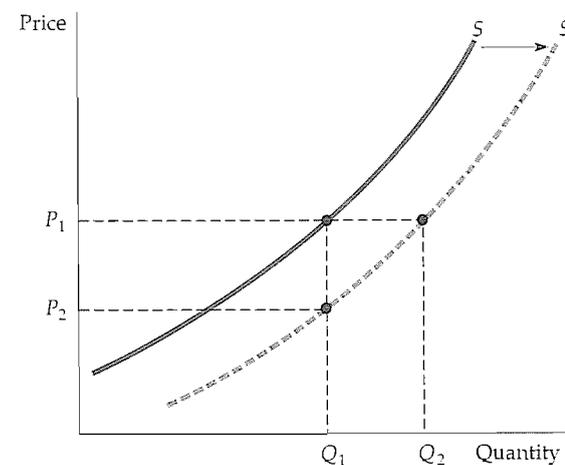


FIGURE 2.1 The Supply Curve

The supply curve, labeled S in the figure, shows how the quantity of a good offered for sale changes as the price of the good changes. The supply curve is upward sloping; the higher the price, the more firms are able and willing to produce and sell. If production costs fall, firms can produce the same quantity at a lower price or a larger quantity at the same price. The supply curve then shifts to the right.

Note that the supply curve slopes upward. In other words, the higher the price, the more that firms are able and willing to produce and sell. For example, a higher price may enable existing firms to expand production by hiring extra workers or by having existing workers work overtime (at greater cost to the firm). Likewise, they may expand production over a longer period of time by increasing the size of their plants. A higher price may also attract new firms to the market. These newcomers face higher costs because of their inexperience in the market and would therefore have found entry uneconomical at a lower price.

Other Variables That Affect Supply The quantity supplied can depend on other variables besides price. For example, the quantity that producers are willing to sell depends not only on the price they receive but also on their production costs, including wages, interest charges, and the costs of raw materials. The supply curve labeled S in Figure 2.1 was drawn for particular values of these other variables. A change in the values of one or more of these variables translates into a shift in the supply curve. Let's see how this might happen.

The supply curve S in Figure 2.1 says that at a price P_1 , the quantity produced and sold would be Q_1 . Now suppose that the cost of raw materials falls. How does this affect the supply curve?

Lower raw material costs—indeed, lower costs of any kind—make production more profitable, encouraging existing firms to expand production and enabling new firms to enter the market. If at the same time the market price stayed constant at P_1 , we would expect to observe a greater quantity supplied. Figure 2.1 shows this as an increase from Q_1 to Q_2 . When production costs decrease, output increases no matter what the market price happens to be. The entire supply curve thus shifts to the right, which is shown in the figure as a shift from S to S' .

Another way of looking at the effect of lower raw material costs is to imagine that the quantity produced stays fixed at Q_1 and then ask what price firms would require to produce this quantity. Because their costs are lower, they would require a lower price— P_2 . This would be the case no matter what quantity was produced. Again, we see in Figure 2.1 that the supply curve must shift to the right.

We have seen that the response of quantity supplied to changes in price can be represented by movements along the supply curve. However, the response of supply to changes in other supply-determining variables is shown graphically as a shift of the supply curve itself. To distinguish between these two graphical depictions of supply changes, economists often use the phrase *change in supply* to refer to shifts in the supply curve, while reserving the phrase *change in the quantity supplied* to apply to movements along the supply curve.

The Demand Curve

The **demand curve** shows how much of a good consumers are willing to buy as the price per unit changes. We can write this relationship between quantity demanded and price as an equation:

$$Q_D = Q_D(P)$$

or we can draw it graphically, as in Figure 2.2. Note that the demand curve in that figure, labeled D , slopes downward: Consumers are usually ready to buy more if the price is lower. For example, a lower price may encourage consumers who have already been buying the good to consume larger quantities. Likewise, it may allow other consumers who were previously unable to afford the good to begin buying it.

demand curve Relationship between the quantity of a good that consumers are willing to buy and the price of the good.

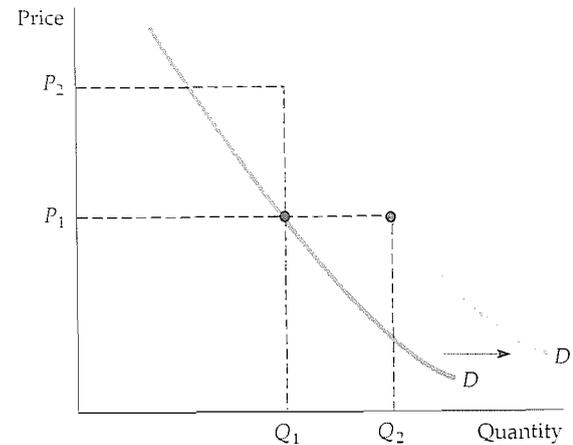


FIGURE 2.2 The Demand Curve

The demand curve, labeled D , shows how the quantity of a good demanded by consumers depends on its price. The demand curve is downward sloping; holding other things equal, consumers will want to purchase more of a good the lower is its price. The quantity demanded may also depend on other variables, such as income, the weather, and the prices of other goods. For most products, the quantity demanded increases when income rises. A higher income level shifts the demand curve to the right.

Of course the quantity of a good that consumers are willing to buy can depend on other things besides its price. *Income* is especially important. With greater incomes, consumers can spend more money on any good, and some consumers will do so for most goods.

Shifting the Demand Curve Let's see what happens to the demand curve if income levels increase. As you can see in Figure 2.2, if the market price were held constant at P_1 , we would expect to see an increase in the quantity demanded—say, from Q_1 to Q_2 , as a result of consumers' higher incomes. Because this increase would occur no matter what the market price, the result would be a *shift to the right of the entire demand curve*. In the figure, this is shown as a shift from D to D' . Alternatively, we can ask what price consumers would pay to purchase a given quantity Q_1 . With greater income, they should be willing to pay a higher price—say, P_2 instead of P_1 in Figure 2.2. Again, *the demand curve will shift to the right*. As we did with supply, we will use the phrase *change in demand* to refer to shifts in the demand curve, and reserve the phrase *change in the quantity demanded* to apply to movements along the demand curve.¹

Substitute and Complementary Goods Changes in the prices of related goods also affect demand. Goods are **substitutes** when an increase in the price of one leads to an increase in the quantity demanded of the other. For example, copper and aluminum are substitute goods. Because one can often be substituted for the other in industrial use, *the quantity of copper demanded will increase if*

substitutes Two goods for which an increase in the price of one leads to an increase in the quantity demanded of the other.

¹ Mathematically, we can write the demand curve as

$$Q_D = D(P, I)$$

where I is disposable income. When we draw a demand curve, we are keeping I fixed.

the price of aluminum increases. Likewise, beef and chicken are substitute goods because most consumers are willing to shift their purchases from one to the other when prices change.

Goods are **complements** when an increase in the price of one leads to a decrease in the quantity demanded of the other. For example, automobiles and gasoline are complementary goods. Because they tend to be used together, a decrease in the price of gasoline increases the quantity demanded for automobiles. Likewise, computers and computer software are complementary goods. The price of computers has dropped dramatically over the past decade, fueling an increase not only in purchases of computers, but also purchases of software packages.

We attributed the shift to the right of the demand curve in Figure 2.2 to an increase in income. However, this shift could also have resulted from either an increase in the price of a substitute good or a decrease in the price of a complementary good. Or it might have resulted from a change in some other variable, such as the weather. For example, demand curves for skis and snowboards will shift to the right when there are heavy snowfalls.

complements Two goods for which an increase in the price of one leads to a decrease in the quantity demanded of the other.

2.2 The Market Mechanism

The next step is to put the supply curve and the demand curve together. We have done this in Figure 2.3. The vertical axis shows the price of a good, P , again measured in dollars per unit. This is now the price that sellers receive for a given quantity supplied, and the price that buyers will pay for a given quantity demanded. The horizontal axis shows the total quantity demanded and supplied, Q , measured in number of units per period.

Equilibrium The two curves intersect at the **equilibrium**, or **market-clearing price** and quantity. At this price (P_0 in Figure 2.3), the quantity supplied and the quantity demanded are just equal (to Q_0). The **market mechanism** is the tendency in a free market for the price to change until the market *clears*—i.e., until

equilibrium (or market-clearing) price Price that equates the quantity supplied to the quantity demanded.

market mechanism Tendency in a free market for price to change until the market clears.

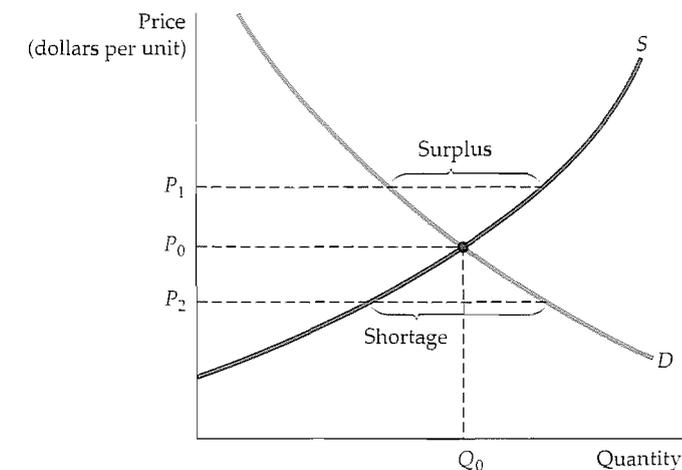


FIGURE 2.3 Supply and Demand

The market clears at price P_0 and quantity Q_0 . At the higher price P_1 , a surplus develops, so price falls. At the lower price P_2 , there is a shortage, so price is bid up.

the quantity supplied and the quantity demanded are equal. At this point, because there is neither excess demand nor excess supply, there is no pressure for the price to change further. Supply and demand might not always be in equilibrium, and some markets might not clear quickly when conditions change suddenly. The *tendency*, however, is for markets to clear.

To understand why markets tend to clear, suppose the price were initially above the market-clearing level—say, P_1 in Figure 2.3. Producers will try to produce and sell more than consumers are willing to buy. A **surplus**—a situation in which the quantity supplied exceeds the quantity demanded—will result. To sell this surplus—or at least to prevent it from growing—producers would begin to lower prices. Eventually, as price fell, quantity demanded would increase, and quantity supplied would decrease until the equilibrium price P_0 was reached.

The opposite would happen if the price were initially below P_0 —say, at P_2 . A **shortage**—a situation in which the quantity demanded exceeds the quantity supplied—would develop, and consumers would be unable to purchase all they would like. This would put upward pressure on price as consumers tried to outbid one another for existing supplies and producers reacted by increasing price and expanding output. Again, the price would eventually reach P_0 .

When Can We Use the Supply-Demand Model? When we draw and use supply and demand curves, we are assuming that at any given price, a given quantity will be produced and sold. This assumption makes sense only if a market is at least roughly *competitive*. By this we mean that both sellers and buyers should have little *market power*—i.e., little ability *individually* to affect the market price.

Suppose instead that supply were controlled by a single producer—a monopolist. In this case, there will no longer be a simple one-to-one relationship between price and the quantity supplied. Why? Because a monopolist's behavior depends on the shape and position of the demand curve. If the demand curve shifts in a particular way, it may be in the monopolist's interest to keep the quantity fixed but change the price, or to keep the price fixed and change the quantity. (How this could occur is explained in Chapter 10.) Thus when we work with supply and demand curves, we implicitly assume that we are referring to a competitive market.

2.3 Changes in Market Equilibrium

We have seen how supply and demand curves shift in response to changes in such variables as wage rates, capital costs, and income. We have also seen how the market mechanism results in an equilibrium in which the quantity supplied equals the quantity demanded. Now we will see how that equilibrium changes in response to shifts in the supply and demand curves.

Let's begin with a shift in the supply curve. In Figure 2.4, the supply curve has shifted from S to S' (as it did in Figure 2.1), perhaps as a result of a decrease in the price of raw materials. As a result, the market price drops (from P_1 to P_3), and the total quantity produced increases (from Q_1 to Q_3). This is what we would expect: Lower costs result in lower prices and increased sales. (Indeed, gradual decreases in costs resulting from technological progress and better management are an important driving force behind economic growth.)

surplus Situation in which the quantity supplied exceeds the quantity demanded.

shortage Situation in which the quantity demanded exceeds the quantity supplied.

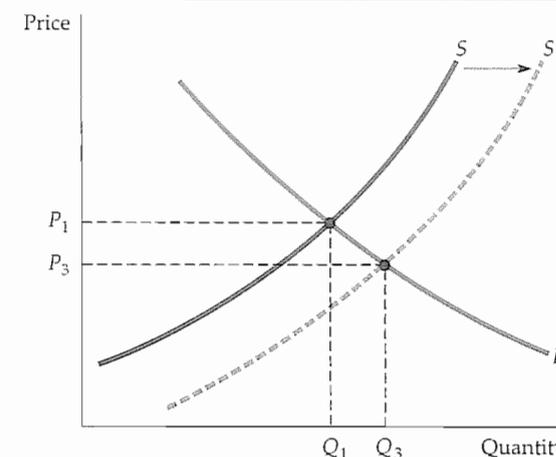


FIGURE 2.4 New Equilibrium Following Shift in Supply

When the supply curve shifts to the right, the market clears at a lower price P_3 and a larger quantity Q_3 .

Figure 2.5 shows what happens following a rightward shift in the demand curve resulting from, say, an increase in income. A new price and quantity result after demand comes into equilibrium with supply. As shown in Figure 2.5, we would expect to see consumers pay a higher price, P_3 , and firms produce a greater quantity, Q_3 , as a result of an increase in income.

In most markets, both the demand and supply curves shift from time to time. Consumers' disposable incomes change as the economy grows (or contracts, during economic recessions). The demands for some goods shift with the seasons (e.g., fuels, bathing suits, umbrellas), with changes in the prices of related goods (an increase in oil prices increases the demand for natural gas), or simply

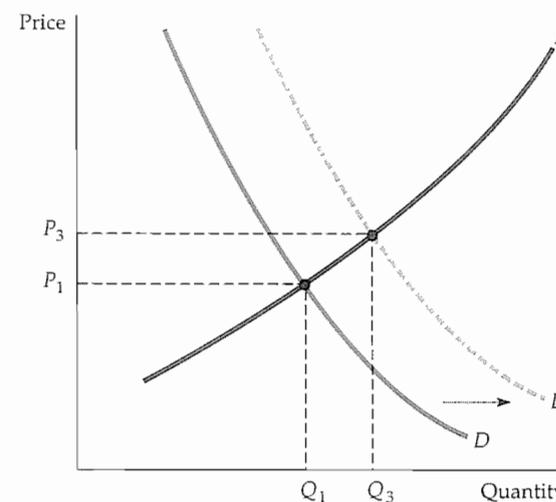


FIGURE 2.5 New Equilibrium Following Shift in Demand

When the demand curve shifts to the right, the market clears at a higher price P_3 and a larger quantity Q_3 .

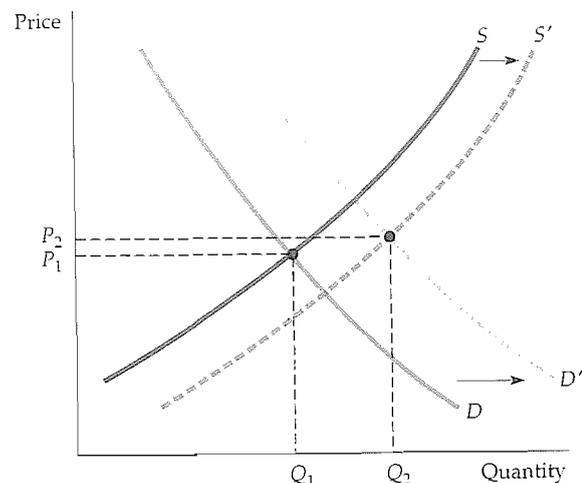


FIGURE 2.6 New Equilibrium Following Shifts in Supply and Demand

Supply and demand curves shift over time as market conditions change. In this example, rightward shifts of the supply and demand curves lead to a slightly higher price and a much larger quantity. In general, changes in price and quantity depend on the amount by which each curve shifts and the shape of each curve.

with changing tastes. Similarly wage rates, capital costs, and the prices of raw materials also change from time to time, and these changes shift the supply curve.

Supply and demand curves can be used to trace the effects of these changes. In Figure 2.6, for example, shifts to the right of both supply and demand result in a slightly higher price (from P_1 to P_2) and a much larger quantity (from Q_1 to Q_2). In general, price and quantity will change depending both on how much the supply and demand curves shift and on the shapes of those curves. To predict the sizes and directions of such changes, we must be able to characterize quantitatively the dependence of supply and demand on price and other variables. We will turn to this task in the next section.

EXAMPLE 2.1 The Price of Eggs and the Price of a College Education Revisited

In Example 1.2, we saw that from 1970 to 1998, the real (constant-dollar) price of eggs fell by 59 percent, while the real price of a college education rose by 81 percent. What caused this large decline in egg prices and large increase in the price of college?

We can understand these price changes by examining the behavior of supply and demand for each good, as shown in Figure 2.7. For eggs, the mechanization of poultry farms sharply reduced the cost of producing eggs, shifting the supply curve downward. At the same time, the demand curve for eggs shifted to the left as a more health-conscious population changed its eating habits and tended to avoid eggs. As a result, the real price of eggs declined sharply, but total annual consumption increased only slightly (from 5300 million dozen to 5500 million dozen).

As for college, supply and demand shifted in the opposite directions. Increases in the costs of equipping and maintaining modern classrooms, labo-

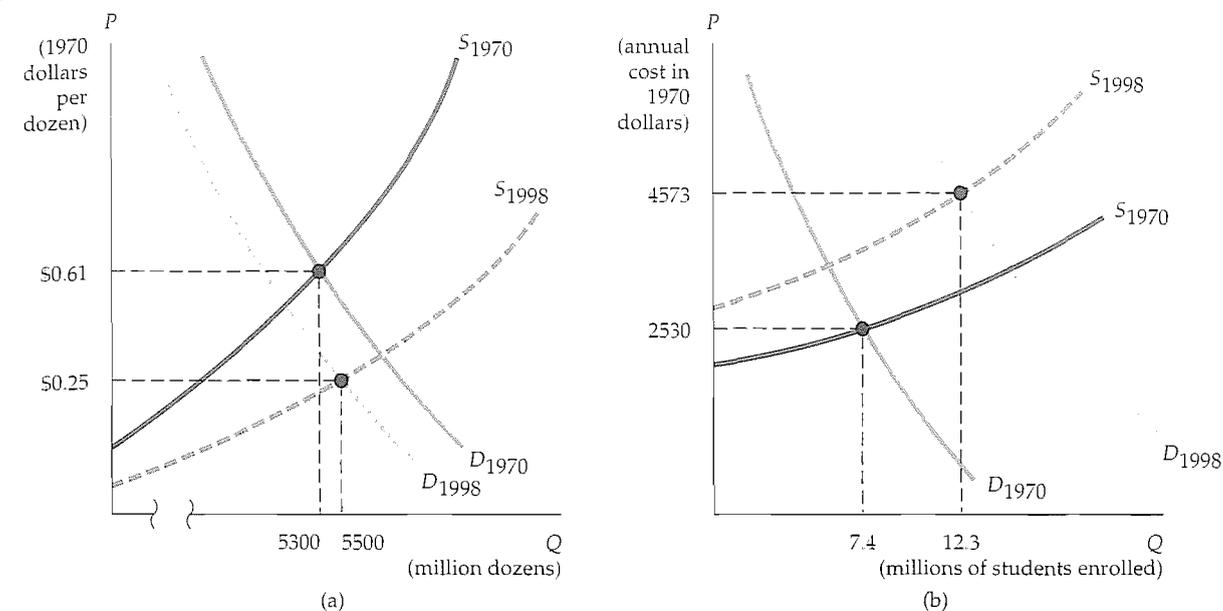


FIGURE 2.7(a) Market for Eggs
(b) Market for College Education

(a) The supply curve for eggs shifted downward as production costs fell; the demand curve shifted to the left as consumer preferences changed. As a result, the real price of eggs fell sharply and egg consumption fell slightly. (b) The supply curve for a college education shifted up as the costs of equipment, maintenance, and staffing rose. The demand curve shifted to the right as a growing number of high school graduates desired a college education. As a result, both price and enrollments rose sharply.

ratories, and libraries, along with increases in faculty salaries, pushed the supply curve up. At the same time, the demand curve shifted to the right as a larger and larger percentage of a growing number of high school graduates decided that a college education was essential. Thus, despite the increase in price, 1998 found more than 12 million students enrolled in undergraduate college degree programs, compared with 7.4 million in 1970.

EXAMPLE 2.2 Wage Inequality in the United States

Although the U.S. economy has grown vigorously over the past two decades, the gains from this growth have not been shared equally by all. Skilled high-income workers have seen their wages grow substantially, while the wages of unskilled low-income workers have, in real terms, actually fallen slightly. Overall, there has been growing inequality in the distribution of earnings, a phenomenon which began around 1980 and has accelerated in recent years. For example, from 1977 to 1999, the top 20 percent of the income distribution experienced an average increase in real (inflation-adjusted) after-tax incomes of more than 40 percent, while the bottom 20 percent of the income distribution *dropped* by over 10 percent. If this increase in inequality continues during the coming decade, it could lead to social unrest and have other troubling implications for American society.

Why has income distribution become so much more unequal during the past two decades? The answer is in the supply and demand for workers. While the supply of unskilled workers—people with limited educations—has grown substantially, the demand for them has risen only slightly. This shift of the supply curve to the right, combined with little movement of the demand curve, has caused wages of unskilled workers to fall. On the other hand, while the supply of skilled workers—e.g., engineers, scientists, managers, and economists—has grown slowly, the demand has risen dramatically, pushing wages up. (We leave it to you as an exercise to draw supply and demand curves and show how they have shifted, as was done in Example 2.1.)

These trends are evident in the behavior of wages for different categories of employment. For example, the real (inflation-adjusted) earnings of managerial and professional workers rose by more than 8 percent from 1983 to 1998. Over the same period, the real incomes of relatively unskilled service workers (such as restaurant workers, sales clerks, and janitorial workers) fell by more than 5 percent.

Most projections point to a continuation of this phenomenon during the beginning of the new millennium. As the high-tech sectors of the American economy grow, the demand for highly skilled workers is likely to increase further. At the same time, the computerization of offices and factories will further reduce the demand for unskilled workers. (This trend is discussed further in Example 14.7.) These changes can only exacerbate wage inequality.

EXAMPLE 2.3 The Long-Run Behavior of Natural Resource Prices

Many people are concerned about the earth's natural resources. At issue is whether our energy and mineral resources are likely to be depleted in the near future, leading to sharp price increases that could bring an end to economic growth. An analysis of supply and demand can give us some perspective.

The earth does indeed have only a finite amount of mineral resources, such as copper, iron, coal, and oil. During the past century, however, the prices of these and most other natural resources have declined or remained roughly constant relative to overall prices. Figure 2.8, for example, shows the price of copper in real terms (adjusted for inflation), together with the quantity consumed from 1880 to 1998. (Both are shown as an index, with 1880 = 1.) Despite short-term variations in price, no significant long-term increase has occurred, even though annual consumption is now about 100 times greater than in 1880. Similar patterns hold for other mineral resources, such as iron, oil, and coal.²

The demands for these resources grew along with the world economy. (These shifts in the demand curve are illustrated in Figure 2.9.) But as demand grew, production costs fell. The decline was due first to the discovery of new and bigger

² The data in Figure 2.8 are from Robert S. Manthey, *Natural Resource Commodities—A Century of Statistics* (Baltimore: Johns Hopkins University Press, 1978), supplemented after 1973 with data from the U.S. Bureau of Mines and from the World Bank.

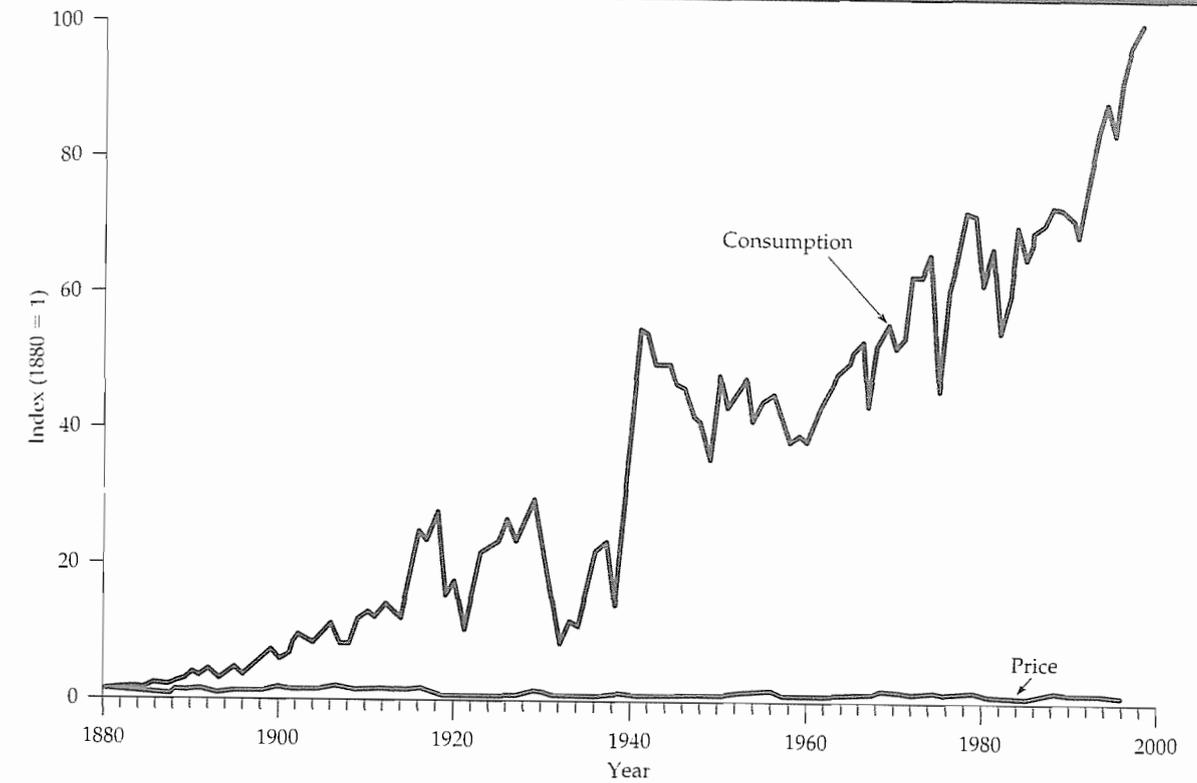


FIGURE 2.8 Consumption and Price of Copper, 1880–1998

Although annual consumption has increased about a hundredfold, the real (inflation-adjusted) price has not changed much.

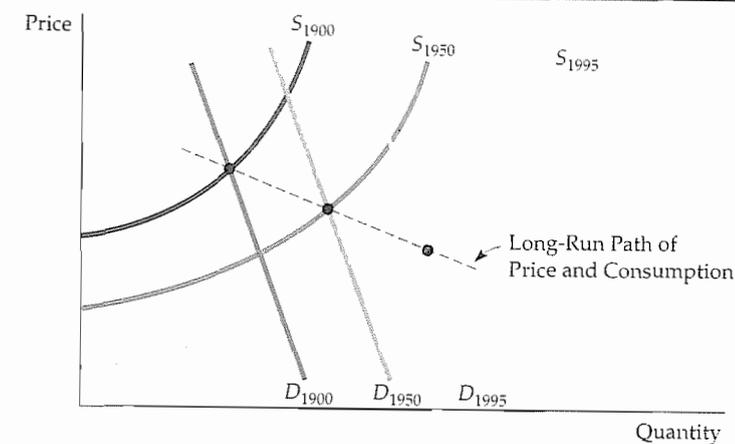


FIGURE 2.9 Long-Run Movements of Supply and Demand for Mineral Resources

Although demand for most resources has increased dramatically over the past century, prices have fallen or risen only slightly in real (inflation-adjusted) terms because cost reductions have shifted the supply curve to the right just as dramatically.

deposits, which were cheaper to mine, and then to technical progress and the economic advantage of mining and refining on a large scale. As a result, the supply curve shifted over time to the right. Over the long term, because increases in supply were greater than increases in demand, price often fell, as shown in Figure 2.9.

This is not to say that the prices of copper, iron, and coal will decline or remain constant forever. After all, these resources are *finite*. But as prices begin to rise, consumption will likely shift, at least in part, to substitute materials. Copper, for example, has already been replaced in many applications by aluminum and, more recently, in electronic applications by fiber optics. (See Example 2.7 for a more detailed discussion of copper prices.)

2.4 Elasticities of Supply and Demand

We have seen that the demand for a good depends not only on its price, but also on consumer income and on the prices of other goods. Likewise, supply depends both on price and on variables that affect production cost. For example, if the price of coffee increases, the quantity demanded will fall, and the quantity supplied will rise. Often, however, we want to know *how much* the quantity supplied or demanded will rise or fall. How sensitive is the demand for coffee to its price? If price increases by 10 percent, how much will the quantity demanded change? How much will it change if income rises by 5 percent? We use *elasticities* to answer questions like these.

An **elasticity** measures the sensitivity of one variable to another. Specifically, it is a number that tells us *the percentage change that will occur in one variable in response to a 1-percent increase in another variable*. For example, the *price elasticity of demand* measures the sensitivity of quantity demanded to price changes. It tells us what the percentage change in the quantity demanded for a good will be following a 1-percent increase in the price of that good.

Price Elasticity of Demand Let's look at this in more detail. Denoting quantity and price by Q and P , we write the **price elasticity of demand** as

$$E_p = (\% \Delta Q) / (\% \Delta P)$$

where $\% \Delta Q$ simply means "percentage change in Q " and $\% \Delta P$ means "percentage change in P ." (The symbol Δ is the Greek capital letter *delta*; it means "the change in." So ΔX means "the change in the variable X ," say, from one year to the next.) The percentage change in a variable is just *the absolute change in the variable divided by the original level of the variable*. (If the Consumer Price Index were 200 at the beginning of the year and increased to 204 by the end of the year, the percentage change—or annual rate of inflation—would be $4/200 = .02$, or 2 percent.) Thus we can also write the price elasticity of demand as follows:³

$$E_p = \frac{\Delta Q/Q}{\Delta P/P} = \frac{P \Delta Q}{Q \Delta P} \quad (2.1)$$

The price elasticity of demand is usually a negative number. When the price of a good increases, the quantity demanded usually falls. Thus $\Delta Q/\Delta P$ (the change in quantity for a change in price) is negative, as is E_p .

³ In terms of infinitesimal changes (letting the ΔP become very small), $E_p = (P/Q)(\Delta Q/\Delta P)$.

elasticity Percentage change in one variable resulting from a 1-percent increase in another.

price elasticity of demand Percentage change in quantity demanded of a good resulting from a 1-percent increase in its price.

When the price elasticity is greater than 1 in magnitude, we say that demand is *price elastic* because the percentage decline in quantity demanded is greater than the percentage increase in price. If the price elasticity is less than 1 in magnitude, demand is said to be *price inelastic*. In general, the price elasticity of demand for a good depends on the availability of other goods that can be substituted for it. When there are close substitutes, a price increase will cause the consumer to buy less of the good and more of the substitute. Demand will then be highly price elastic. When there are no close substitutes, demand will tend to be price inelastic.

Linear Demand Curve Equation (2.1) says that the price elasticity of demand is the change in quantity associated with a change in price ($\Delta Q/\Delta P$) times the ratio of price to quantity (P/Q). But as we move down the demand curve, $\Delta Q/\Delta P$ may change, and the price and quantity will always change. Therefore, the price elasticity of demand must be measured *at a particular point on the demand curve* and will generally change as we move along the curve.

This principle is easiest to see for a **linear demand curve**—that is, a demand curve of the form

$$Q = a - bP$$

As an example, consider the demand curve

$$Q = 8 - 2P$$

For this curve, $\Delta Q/\Delta P$ is constant and equal to -2 (a ΔP of 1 results in a ΔQ of -2). However, the curve does *not* have a constant elasticity. Observe from Figure 2.10 that as we move down the curve, the ratio P/Q falls; the elasticity

linear demand curve Demand curve that is a straight line.

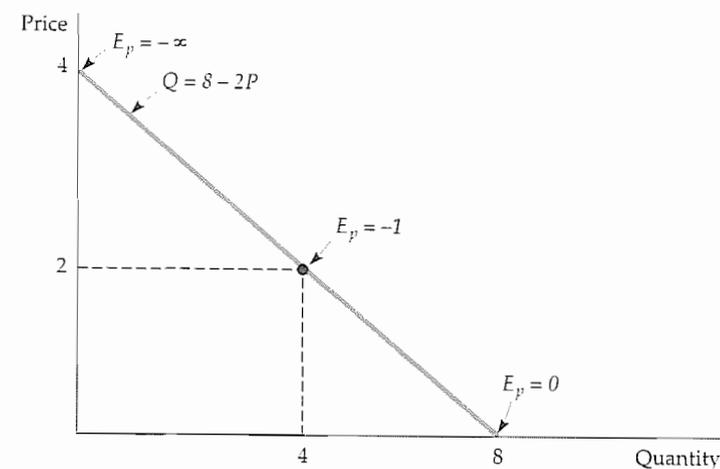


FIGURE 2.10 Linear Demand Curve

The price elasticity of demand depends not only on the slope of the demand curve but also on the price and quantity. The elasticity, therefore, varies along the curve as price and quantity change. Slope is constant for this linear demand curve. Near the top, because price is high and quantity is small, the elasticity is large in magnitude. The elasticity becomes smaller as we move down the curve.

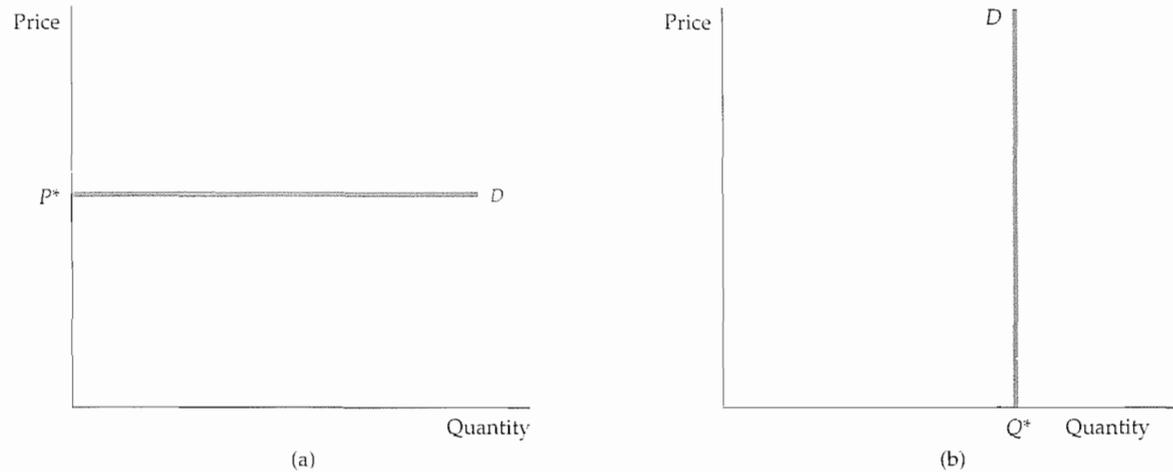


FIGURE 2.11(a) Infinitely Elastic Demand
(b) Completely Inelastic Demand

(a) For a horizontal demand curve, $\Delta Q/\Delta P$ is infinite. Because a tiny change in price leads to an enormous change in demand, the elasticity of demand is infinite. (b) For a vertical demand curve, $\Delta Q/\Delta P$ is zero. Because the quantity demanded is the same no matter what the price, the elasticity of demand is zero.

therefore decreases in magnitude. Near the intersection of the curve with the price axis, Q is very small, so $E_p = -2(P/Q)$ is large in magnitude. When $P = 2$ and $Q = 1$, $E_p = -2$. At the intersection with the quantity axis, $P = 0$ so $E_p = 0$.

Because we draw demand (and supply) curves with price on the vertical axis and quantity on the horizontal axis, $\Delta Q/\Delta P = (1/\text{slope of curve})$. As a result, for any price and quantity combination, the steeper the slope of the curve, the less elastic is demand. Figure 2.11 shows two special cases. Figure 2.11(a) shows a demand curve reflecting **infinitely elastic demand**: Consumers will buy as much as they can at a single price P^* . For even the smallest increase in price above this level, quantity demanded drops to zero, and for any decrease in price, quantity demanded increases without limit. The demand curve in Figure 2.11(b), on the other hand, reflects **completely inelastic demand**: Consumers will buy a fixed quantity Q^* , no matter what the price.

Other Demand Elasticities We will also be interested in elasticities of demand with respect to other variables besides price. For example, demand for most goods usually rises when aggregate income rises. The **income elasticity of demand** is the percentage change in the quantity demanded, Q , resulting from a 1-percent increase in income I :

$$E_I = \frac{\Delta Q/Q}{\Delta I/I} = \frac{I}{Q} \frac{\Delta Q}{\Delta I} \quad (2.2)$$

The demand for some goods is also affected by the prices of other goods. For example, because butter and margarine can easily be substituted for each other, the demand for each depends on the price of the other. A **cross-price elasticity of demand** refers to the percentage change in the quantity demanded for a good that results from a 1-percent increase in the price of another good. So the

elasticity of demand for butter with respect to the price of margarine would be written as

$$E_{Q_b, P_m} = \frac{\Delta Q_b/Q_b}{\Delta P_m/P_m} = \frac{P_m}{Q_b} \frac{\Delta Q_b}{\Delta P_m} \quad (2.3)$$

where Q_b is the quantity of butter and P_m is the price of margarine.

In this example, the cross-price elasticities will be positive because the goods are *substitutes*: Because they compete in the market, a rise in the price of margarine, which makes butter cheaper relative to margarine, leads to an increase in the quantity of butter demanded. (Because the demand curve for butter will shift to the right, the price of butter will rise.) But this is not always the case. Some goods are *complements*: Because they tend to be used together, an increase in the price of one tends to push down the consumption of the other. Gasoline and motor oil are an example. If the price of gasoline goes up, the quantity of gasoline demanded falls—motorists will drive less. But the demand for motor oil also falls. (The entire demand curve for motor oil shifts to the left.) Thus, the cross-price elasticity of motor oil with respect to gasoline is negative.

Elasticities of Supply Elasticities of supply are defined in a similar manner. The **price elasticity of supply** is the percentage change in the quantity supplied resulting from a 1-percent increase in price. This elasticity is usually positive because a higher price gives producers an incentive to increase output.

We can also refer to elasticities of supply with respect to such variables as interest rates, wage rates, and the prices of raw materials and other intermediate goods used to manufacture the product in question. For example, for most manufactured goods, the elasticities of supply with respect to the prices of raw materials are negative. An increase in the price of a raw material input means higher costs for the firm; other things being equal, therefore, the quantity supplied will fall.

price elasticity of supply
 Percentage change in quantity supplied resulting from a 1-percent increase in price.

EXAMPLE 2.4 The Market for Wheat

Wheat is an important agricultural commodity, and the wheat market has been studied extensively by agricultural economists. During the 1980s and 1990s, changes in the wheat market had major implications for both American farmers and U.S. agricultural policy. To understand what happened, let's examine the behavior of supply and demand over this period.

From statistical studies, we know that for 1981 the supply curve for wheat was approximately as follows:⁴

$$\text{Supply: } Q_S = 1800 + 240P$$

⁴ For a survey of statistical studies of the demand and supply of wheat and an analysis of evolving market conditions, see Larry Salathe and Sudchada Langley, "An Empirical Analysis of Alternative Export Subsidy Programs for U.S. Wheat," *Agricultural Economics Research* 38, No. 1 (Winter 1986). The supply and demand curves in this example are based on the studies they survey.

infinitely elastic demand
 Consumers will buy as much of a good as they can get at a single price, but for any higher price the quantity demanded drops to zero, while for any lower price the quantity demanded increases without limit.

completely inelastic demand
 Consumers will buy a fixed quantity of a good regardless of its price.

income elasticity of demand
 Percentage change in the quantity demanded resulting from a 1-percent increase in income.

cross-price elasticity of demand
 Percentage change in the quantity demanded of one good resulting from a 1-percent increase in the price of another.

where price is measured in nominal dollars per bushel and quantities are in millions of bushels per year. These studies also indicate that in 1981 the demand curve for wheat was

$$\text{Demand: } Q_D = 3550 - 266P$$

By setting the quantity supplied equal to the quantity demanded, we can determine the market-clearing price of wheat for 1981:

$$\begin{aligned} Q_S &= Q_D \\ 1800 + 240P &= 3550 - 266P \\ 506P &= 1750 \\ P &= \$3.46 \text{ per bushel} \end{aligned}$$

To find the market-clearing quantity, substitute this price of \$3.46 into either the supply curve equation or the demand curve equation. Substituting into the supply curve equation, we get

$$Q = 1800 + (240)(3.46) = 2630 \text{ million bushels}$$

What are the price elasticities of demand and supply at this price and quantity? We use the demand curve to find the price elasticity of demand:

$$E_P^D = \frac{P}{Q} \frac{\Delta Q_D}{\Delta P} = \frac{3.46}{2630} (-266) = -0.35$$

Thus demand is inelastic. We can likewise calculate the price elasticity of supply:

$$E_P^S = \frac{P}{Q} \frac{\Delta Q_S}{\Delta P} = \frac{3.46}{2630} (240) = 0.32$$

Because these supply and demand curves are linear, the price elasticities will vary as we move along the curves. For example, suppose that a drought caused the supply curve to shift far enough to the left to push the price up to \$4.00 per bushel. In this case the quantity demanded would fall to $3550 - (266)(4.00) = 2486$ million bushels. At this price and quantity, the elasticity of demand would be

$$E_P^D = \frac{4.00}{2486} (-266) = -0.43$$

The wheat market has evolved over the years, in part because of changes in the demand for wheat. The demand for wheat has two components: domestic demand (demand by U.S. consumers) and export demand (demand by foreign consumers). During the 1980s and 1990s, domestic demand for wheat rose only slightly (due to modest increases in population and income). Export demand, however, fell sharply. There were several reasons. First and foremost was the success of the Green Revolution in agriculture: Developing countries like India, which had been large importers of wheat, became increasingly self-sufficient. In addition, European countries adopted protectionist policies that subsidized their own production and imposed tariff barriers against imported wheat.

In 1998, demand and supply were

$$\text{Demand: } Q_D = 3244 - 283P$$

$$\text{Supply: } Q_S = 1944 + 207P$$

Once again, equating quantity supplied and quantity demanded yields the market-clearing (nominal) price and quantity:

$$\begin{aligned} 1944 + 207P &= 3244 - 283P \\ P &= \$2.65 \text{ per bushel} \end{aligned}$$

$$Q = 3244 - (283)(2.65) = 2494 \text{ million bushels}$$

Thus the price of wheat fell even in nominal terms. (You can check to see that at this price and quantity, the price elasticity of demand was -0.30 and the price elasticity of supply was 0.22 .)

The price of wheat was actually greater than \$3.46 in 1981 because the U.S. government bought wheat through its price-support program. In addition, throughout the 1980s and 1990s, farmers received direct subsidies for the wheat they produced. We discuss how such agricultural policies work and evaluate the costs and benefits for consumers, farmers, and the federal budget in Chapter 9.

2.5 Short-Run versus Long-Run Elasticities

When analyzing demand and supply, it is important to distinguish between the short run and the long run. In other words, if we ask how much demand or supply changes in response to a change in price, we must be clear about *how much time is allowed to pass before measuring the changes in the quantity demanded or supplied*. If we allow only a short time to pass—say, one year or less—then we are dealing with the *short run*. When we refer to the *long run*, we mean that enough time is allowed for consumers or producers to *fully adjust* to the price change. In general, short-run demand and supply curves look very different from their long-run counterparts.

Demand

For many goods, demand is much more price elastic in the long run than in the short run. For one thing, it takes time for people to change their consumption habits. For example, even if the price of coffee rises sharply, the quantity demanded will fall only gradually as consumers begin to drink less. In addition, the demand for a good might be linked to the stock of another good that changes only slowly. For example, the demand for gasoline is much more elastic in the long run than in the short run. A sharply higher price of gasoline reduces the quantity demanded in the short run by causing motorists to drive less, but it has its greatest impact on demand by inducing consumers to buy smaller and more fuel-efficient cars. But because the stock of cars changes only slowly, the quantity of gasoline demanded falls only slowly. Figure 2.12 shows short-run and long-run demand curves for goods such as these.

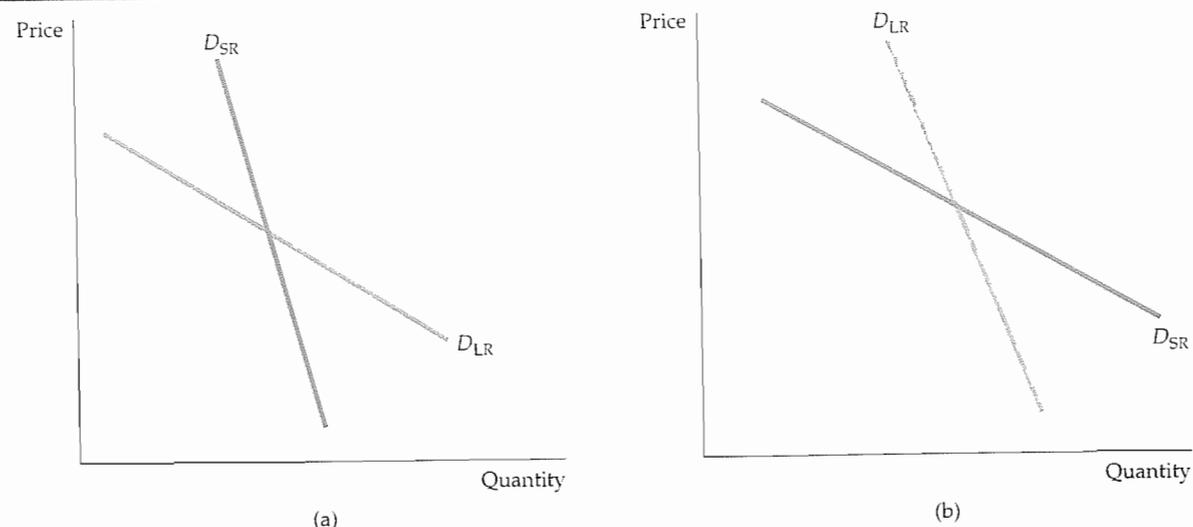


FIGURE 2.12(a) Gasoline: Short-Run and Long-Run Demand Curves
(b) Automobiles: Short-Run and Long-Run Demand Curves

(a) In the short run, an increase in price has only a small effect on the quantity of gasoline demanded. Motorists may drive less, but they will not change the kinds of cars they are driving overnight. In the longer run, however, because they will shift to smaller and more fuel-efficient cars, the effect of the price increase will be larger. Demand, therefore, is more elastic in the long run than in the short run. (b) The opposite is true for automobile demand. If price increases, consumers initially defer buying new cars; thus annual quantity demanded falls sharply. In the longer run, however, old cars wear out and must be replaced; thus annual quantity demanded picks up. Demand, therefore, is less elastic in the long run than in the short run.

Demand and Durability On the other hand, for some goods just the opposite is true—demand is more elastic in the short run than in the long run. Because these goods (automobiles, refrigerators, televisions, or the capital equipment purchased by industry) are *durable*, the total stock of each good owned by consumers is large relative to annual production. As a result, a small change in the total stock that consumers want to hold can result in a large percentage change in the level of purchases.

Suppose, for example, that the price of refrigerators goes up 10 percent, causing the total stock of refrigerators that consumers want to hold to drop 5 percent. Initially, this will cause purchases of new refrigerators to drop much more than 5 percent. But eventually, as consumers' refrigerators depreciate (and units must be replaced), the quantity demanded will increase again. In the long run the total stock of refrigerators owned by consumers will be about 5 percent less than before the price increase. In this case, while the long-run price elasticity of demand for refrigerators would be $-.05/.10 = -0.5$, the short-run elasticity would be much larger in magnitude.

Or consider automobiles. Although annual U.S. demand—new car purchases—is about 8 to 11 million, the stock of cars that people own is around 120 million. If automobile prices rise, many people will delay buying new cars. The quantity demanded will fall sharply, even though the total stock of cars that consumers might want to own at these higher prices falls only a small amount. Eventually, however, because old cars wear out and must be replaced, the quantity of new cars demanded picks up again. As a result, the long-run change in the quantity demanded is much smaller than the short-run change. Figure 2.12(b) shows demand curves for a durable good like automobiles.

Income Elasticities Income elasticities also differ from the short run to the long run. For most goods and services—foods, beverages, fuel, entertainment, etc.—the income elasticity of demand is larger in the long run than in the short run. Consider the behavior of gasoline consumption during a period of strong economic growth during which aggregate income rises by 10 percent. Eventually people will increase gasoline consumption because they can afford to take more trips and perhaps own larger cars. But this change in consumption takes time, and demand initially increases only by a small amount. Thus, the long-run elasticity will be larger than the short-run elasticity.

For a durable good, the opposite is true. Again, consider automobiles. If aggregate income rises by 10 percent, the total stock of cars that consumers will want to own will also rise—say, by 5 percent. But this change means a much larger increase in *current purchases* of cars. (If the stock is 120 million, a 5-percent increase is 6 million, which might be about 60 percent of normal demand in a single year.) Eventually consumers succeed in increasing the total number of cars owned; after the stock has been rebuilt, new purchases are made largely to replace old cars. (These new purchases will still be greater than before because a larger stock of cars outstanding means that more cars need to be replaced each year.) Clearly, the short-run income elasticity of demand will be much larger than the long-run elasticity.

Cyclical Industries Because the demands for durable goods fluctuate so sharply in response to short-run changes in income, the industries that produce

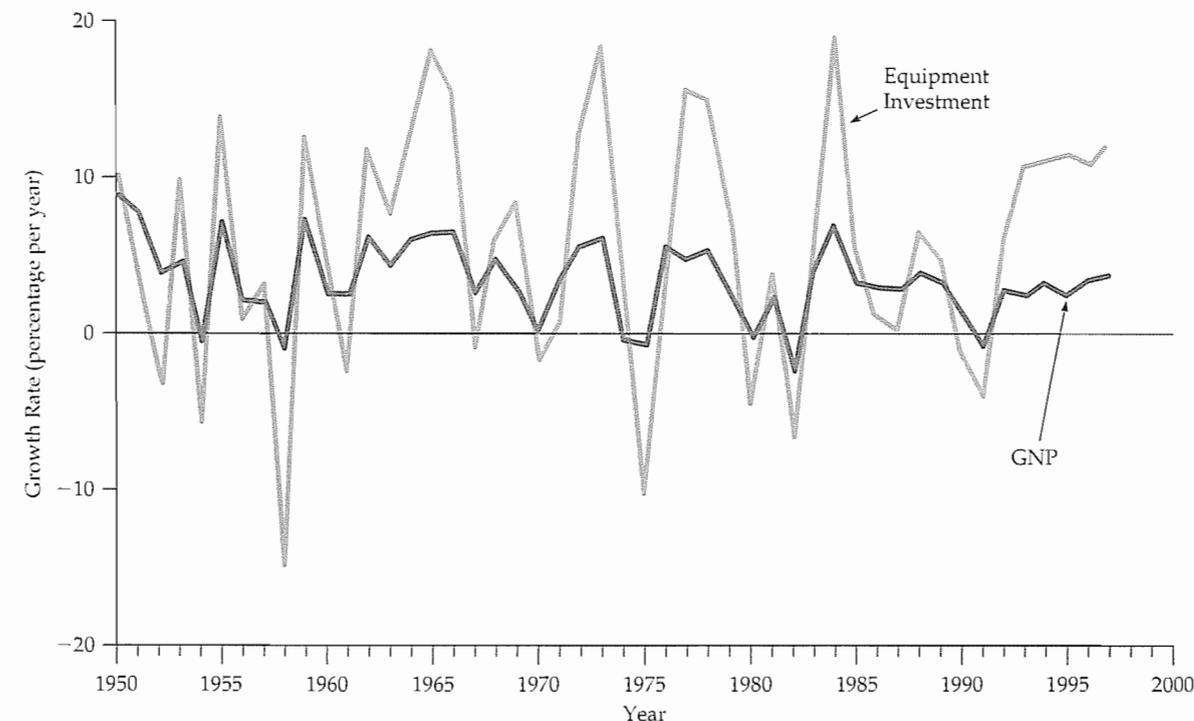


FIGURE 2.13 GNP and Investment in Durable Equipment

Annual growth rates are compared for GNP and investment in durable equipment. Because the short-run GNP elasticity of demand is larger than the long-run elasticity for long-lived capital equipment, changes in investment in equipment magnify changes in GNP. Thus capital goods industries are considered "cyclical."

cyclical industries Industries in which sales tend to magnify cyclical changes in gross national product and national income.

these goods are quite vulnerable to changing macroeconomic conditions, and in particular to the business cycle—recessions and booms. Hence, these industries are often called **cyclical industries**—their sales patterns tend to magnify cyclical changes in gross national product (GNP) and national income.

Figures 2.13 and 2.14 illustrate this principle. Figure 2.13 plots two variables over time: the annual real (inflation-adjusted) rate of growth of GNP and the annual real rate of growth of investment in producers' durable equipment (i.e., machinery and other equipment purchased by firms). Note that although the durable equipment series follows the same pattern as the GNP series, the changes in GNP are magnified. For example, in 1961–1966 GNP grew by at least 4 percent each year. Purchases of durable equipment also grew, but by much more (over 10 percent in 1963–1966). Equipment investment likewise grew much faster than GNP during 1993–1998. On the other hand, during the recessions of 1974–1975, 1982, and 1991, equipment purchases fell by much more than GNP.

Figure 2.14 also shows the real rate of growth of GNP, along with the annual real rates of growth of spending by consumers on durable goods (automobiles, appliances, etc.), and nondurable goods (food, fuel, clothing, etc.). Note that while both consumption series follow GNP, only the durable goods series tends to magnify changes in GNP. Changes in consumption of nondurables are

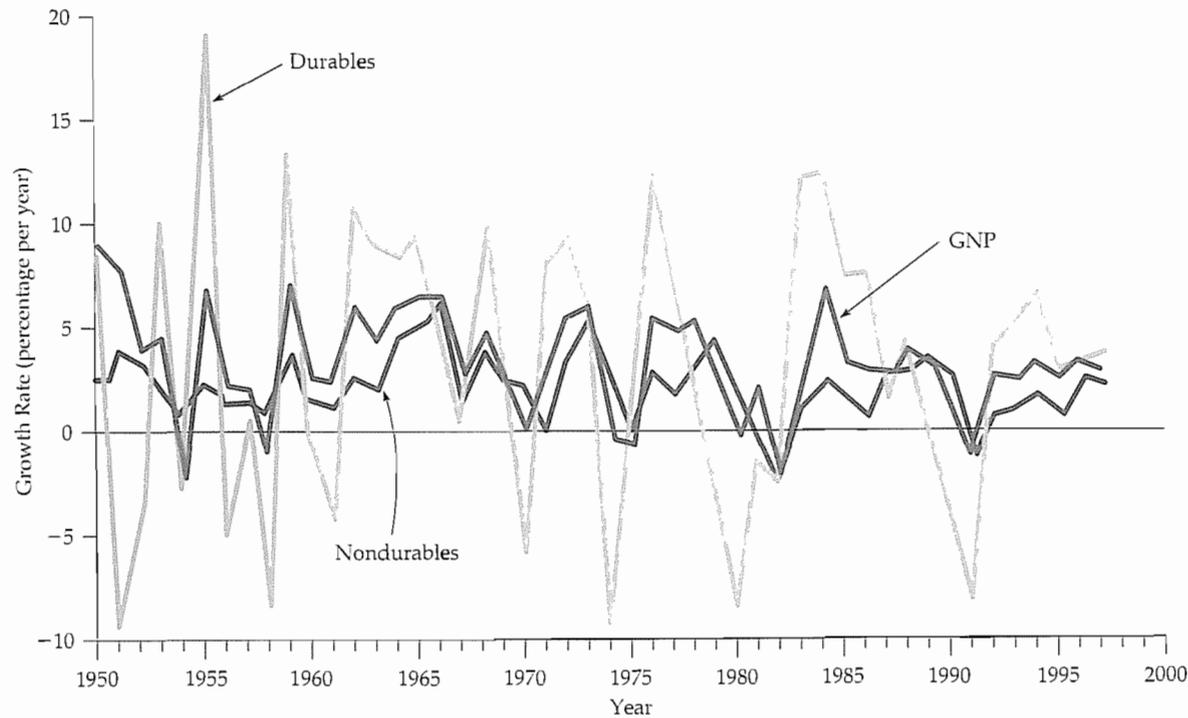


FIGURE 2.14 Consumption of Durables versus Nondurables

Annual growth rates are compared for GNP, consumer expenditures on durable goods (automobiles, appliances, furniture, etc.), and consumer expenditures on nondurable goods (food, clothing, services, etc.). Because the stock of durables is large compared with annual demand, short-run demand elasticities are larger than long-run elasticities. Like capital equipment, industries that produce consumer durables are “cyclical” (i.e., changes in GNP are magnified). This is not true for producers of nondurables.

roughly the same as changes in GNP, but changes in consumption of durables are usually several times larger. This is why companies such as General Motors and General Electric are considered “cyclical”: Sales of cars and electrical appliances are strongly affected by changing macroeconomic conditions.

EXAMPLE 2.5 The Demand for Gasoline and Automobiles

Gasoline and automobiles exemplify some of the different characteristics of demand discussed above. They are complementary goods—an increase in the price of one tends to reduce the demand for the other. In addition, their respective dynamic behaviors (long-run versus short-run elasticities) are just the opposite from each other. For gasoline, the long-run price and income elasticities are larger than the short-run elasticities; for automobiles, the reverse is true.

There have been a number of statistical studies of the demands for gasoline and automobiles. Here we report elasticity estimates from a study that emphasizes the dynamic response of demand.⁵ Table 2.1 shows price and income elasticities of demand for gasoline in the United States for the short run, the long run, and just about everything in between.

Note the large differences between the long-run and the short-run elasticities. Following the sharp increases that occurred in the price of gasoline with the rise of the OPEC oil cartel in 1974, many people (including executives in the automobile and oil industries) claimed that the quantity of gasoline demanded would not change much—that demand was not very elastic. Indeed, for the first year after the price rise, they were right. But demand did eventually change. It just took time for people to alter their driving habits and to replace large cars with smaller and more fuel-efficient ones. This response continued after the second sharp increase in oil prices that occurred in 1979–1980. It was partly because of this response that OPEC could not maintain oil prices above \$30 per barrel, and prices fell.

Table 2.2 shows price and income elasticities of demand for automobiles. Note that the short-run elasticities are much larger than the long-run elasticities. It should be clear from the income elasticities why the automobile industry

TABLE 2.1 Demand for Gasoline

Elasticity	NUMBER OF YEARS ALLOWED TO PASS FOLLOWING A PRICE OR INCOME CHANGE					
	1	2	3	5	10	20
Price	-0.11	-0.22	-0.32	-0.49	-0.82	-1.17
Income	0.07	0.13	0.20	0.32	0.54	0.78

⁵ The elasticity estimates are from R. S. Pindyck, *The Structure of World Energy Demand* (Cambridge, MA: MIT Press, 1979). For related demand studies and elasticity estimates, see Carol Dahl and Thomas Sterner, “Analyzing Gasoline Demand Elasticities: A Survey,” *Energy Economics* (July 1991); Molly Espey, “Watching the Fuel Gauge: An International Model of Automobile Fuel Economy,” *Energy Economics* (April 1996); and David L. Greene, James R. Kahn, and Robert C. Gibson, “Fuel Economy Rebound Effects for U.S. Household Vehicles,” *The Energy Journal* 20, No. 3 (1999).

TABLE 2.2 Demand for Automobiles

Elasticity	NUMBER OF YEARS ALLOWED TO PASS FOLLOWING A PRICE OR INCOME CHANGE					
	1	2	3	5	10	20
Price	-1.20	-0.93	-0.75	-0.55	-0.42	-0.40
Income	3.00	2.33	1.88	1.38	1.02	1.00

is so highly cyclical. For example, GNP fell by nearly 3 percent in real (inflation-adjusted) terms during the 1982 recession, but automobile sales fell by about 8 percent in real terms.⁶ Auto sales recovered, however, during 1983–1985. They also fell by about 8 percent during the 1991 recession (when GNP fell 2 percent), but began to recover in 1993, and rose sharply during 1995–1999.

Supply

Elasticities of supply also differ from the long run to the short run. For most products, long-run supply is much more price elastic than short-run supply: Firms face *capacity constraints* in the short run and need time to expand capacity by building new production facilities and hiring workers to staff them. This is not to say that the quantity supplied will not increase in the short run if price goes up sharply. Even in the short run, firms can increase output by using their existing facilities for more hours per week, paying workers to work overtime, and hiring some new workers immediately. But firms will be able to expand output much more when they have the time to expand their facilities and hire a larger permanent workforce.

For some goods and services, short-run supply is completely inelastic. Rental housing in most cities is an example. In the very short run, there is only a fixed number of rental units. Thus an increase in demand only pushes rents up. In the longer run, and without rent controls, higher rents provide an incentive to renovate existing buildings and construct new ones. As a result, the quantity supplied increases.

For most goods, however, firms can find ways to increase output even in the short run—if the price incentive is strong enough. However, because various constraints make it costly to increase output rapidly, it may require large price increases to elicit small short-run increases in the quantity supplied. We discuss these characteristics of supply in more detail in Chapter 8.

Supply and Durability For some goods, supply is more elastic in the short run than in the long run. Such goods are durable and can be recycled as part of supply if price goes up. An example is the *secondary supply* of metals: the supply from *scrap metal*, which is often melted down and refabricated. When the price of copper goes up, it increases the incentive to convert scrap copper into new supply, so that, initially, secondary supply increases sharply. Eventually, however, the stock of good-quality scrap falls, making the melting, purifying, and refabricating more costly. Secondary supply then contracts. Thus the long-run price elasticity of secondary supply is smaller than the short-run elasticity.

⁶ This includes imports, which were capturing a growing share of the U.S. market. Domestic auto sales fell by even more.

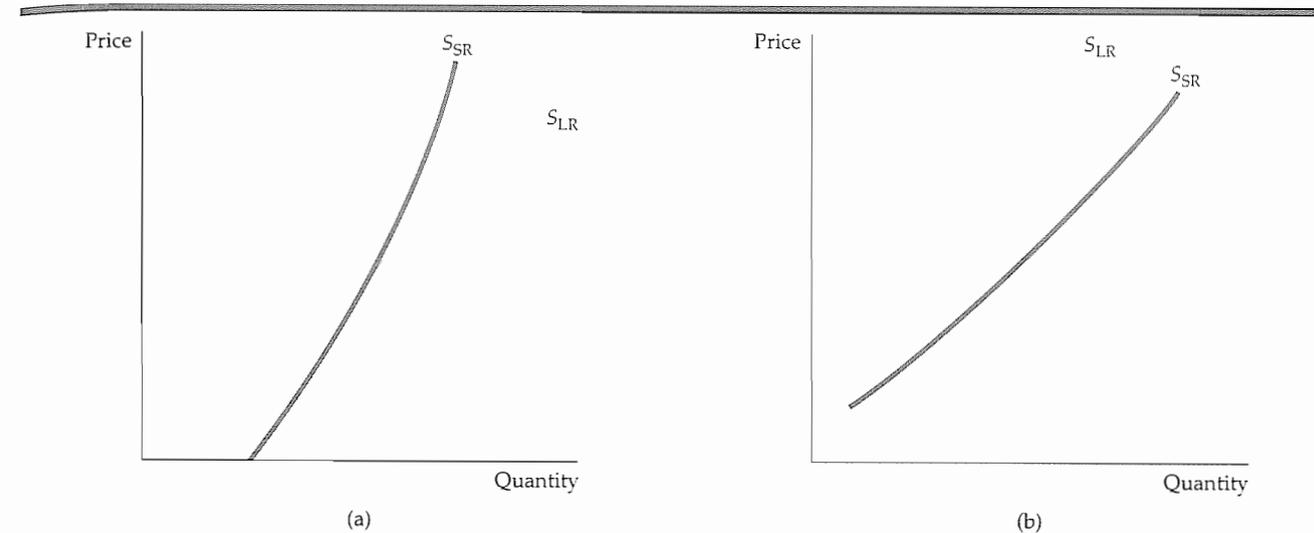


FIGURE 2.15 Copper: Short-Run and Long-Run Supply Curves

Like that of most goods, the supply of primary copper, shown in part (a), is more elastic in the long run. If price increases, firms would like to produce more but are limited by capacity constraints in the short run. In the longer run, they can add to capacity and produce more. Part (b) shows supply curves for secondary copper. If the price increases, there is a greater incentive to convert scrap copper into new supply. Initially, therefore, secondary supply (i.e., supply from scrap) increases sharply. But later, as the stock of scrap falls, secondary supply contracts. Secondary supply is therefore less elastic in the long run than in the short run.

TABLE 2.3 Supply of Copper

PRICE ELASTICITY OF:	SHORT-RUN	LONG-RUN
Primary supply	0.20	1.60
Secondary supply	0.43	0.31
Total supply	0.25	1.50

Figures 2.15(a) and 2.15(b) show short-run and long-run supply curves for primary (production from the mining and smelting of ore) and secondary copper production. Table 2.3 shows estimates of the elasticities for each component of supply and for total supply, based on a weighted average of the component elasticities.⁷ Because secondary supply is only about 20 percent of total supply, the price elasticity of total supply is larger in the long run than in the short run.

EXAMPLE 2.6 The Weather in Brazil and the Price of Coffee in New York

Droughts or subfreezing weather occasionally destroy or damage many of Brazil’s coffee trees. Because Brazil produces much of the world’s coffee, the result is a decrease in the supply of coffee and a sharp run-up in its price.

⁷ These estimates were obtained by aggregating the regional estimates reported in Franklin M. Fisher, Paul H. Cootner, and Martin N. Baily, “An Econometric Model of the World Copper Industry,” *Bell Journal of Economics* 3 (Autumn 1972): 568–609.

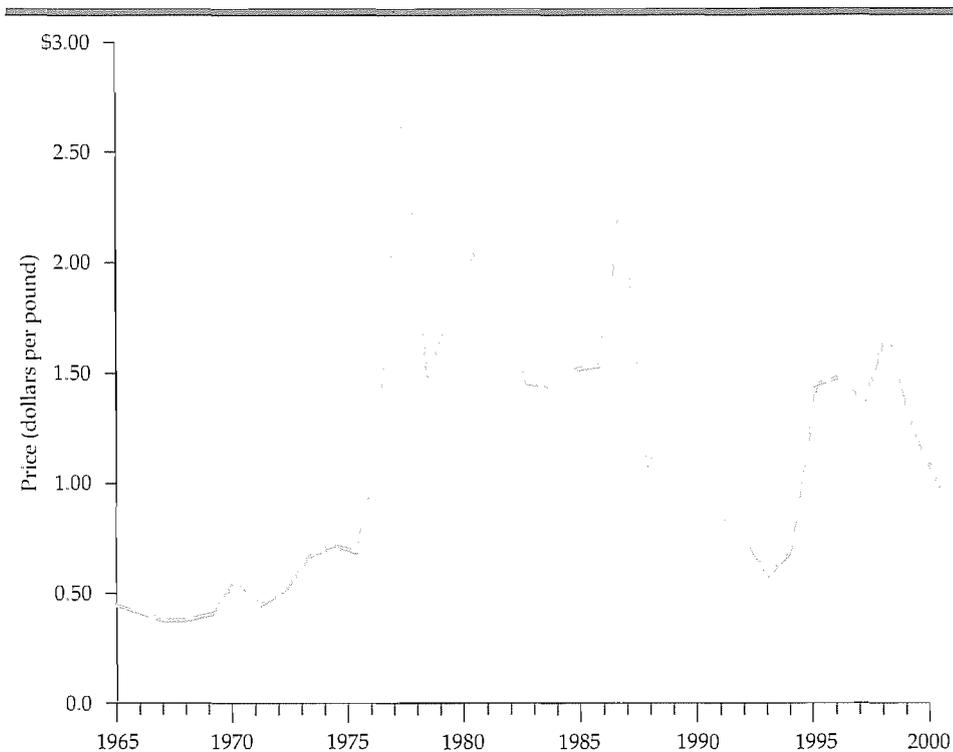


FIGURE 2.16 Price of Brazilian Coffee

When droughts or freezes damage Brazil's coffee trees, the price of coffee can soar. The price usually falls again after a few years, as demand and supply adjust.

In July 1975, for example, a frost destroyed most of Brazil's 1976–1977 coffee crop. (Remember that it is winter in Brazil when it is summer in the northern hemisphere.) As Figure 2.16 shows, the price of a pound of coffee in New York went from 68 cents in 1975 to \$1.23 in 1976 and \$2.70 in 1977. Prices fell, but then jumped again in 1986, after a seven-month drought in 1985 ruined much of Brazil's crop. Finally, starting in June 1994, freezing weather followed by a drought destroyed nearly half of Brazil's 1995–1996 crop. As a result, the price of coffee in 1994–1995 was about double its 1993 level. By 1998, however, the price had dropped considerably.

The run-up price following a freeze or drought is usually short-lived, however. Within a year, price begins to fall; within three or four years, it returns to its earlier levels. In 1978, for example, the price of coffee in New York fell to \$1.48 per pound, and by 1983 it had fallen in real (inflation-adjusted) terms to within a few cents of its prefreeze 1975 price.⁸ Likewise, in 1987 the price of coffee fell to below its predrought 1984 level, and then continued declining until the 1994 freeze.

⁸ During 1980, however, prices temporarily went just above \$2.00 per pound as a result of export quotas imposed under the International Coffee Agreement (ICA). The ICA is essentially a cartel agreement implemented by the coffee-producing countries in 1968. It has been largely ineffective and has seldom had an effect on the price. We discuss cartel pricing in detail in Chapter 12.

Coffee prices behave this way because both demand and supply (especially supply) are much more elastic in the long run than in the short run. Figure 2.17 illustrates this. Note from part (a) of the figure that in the very short run (within one or two months after a freeze), supply is completely inelastic: There are simply a fixed number of coffee beans, some of which have been damaged by the frost. Demand is also relatively inelastic. As a result of the frost, the supply curve shifts to the left, and price increases sharply, from P_0 to P_1 .

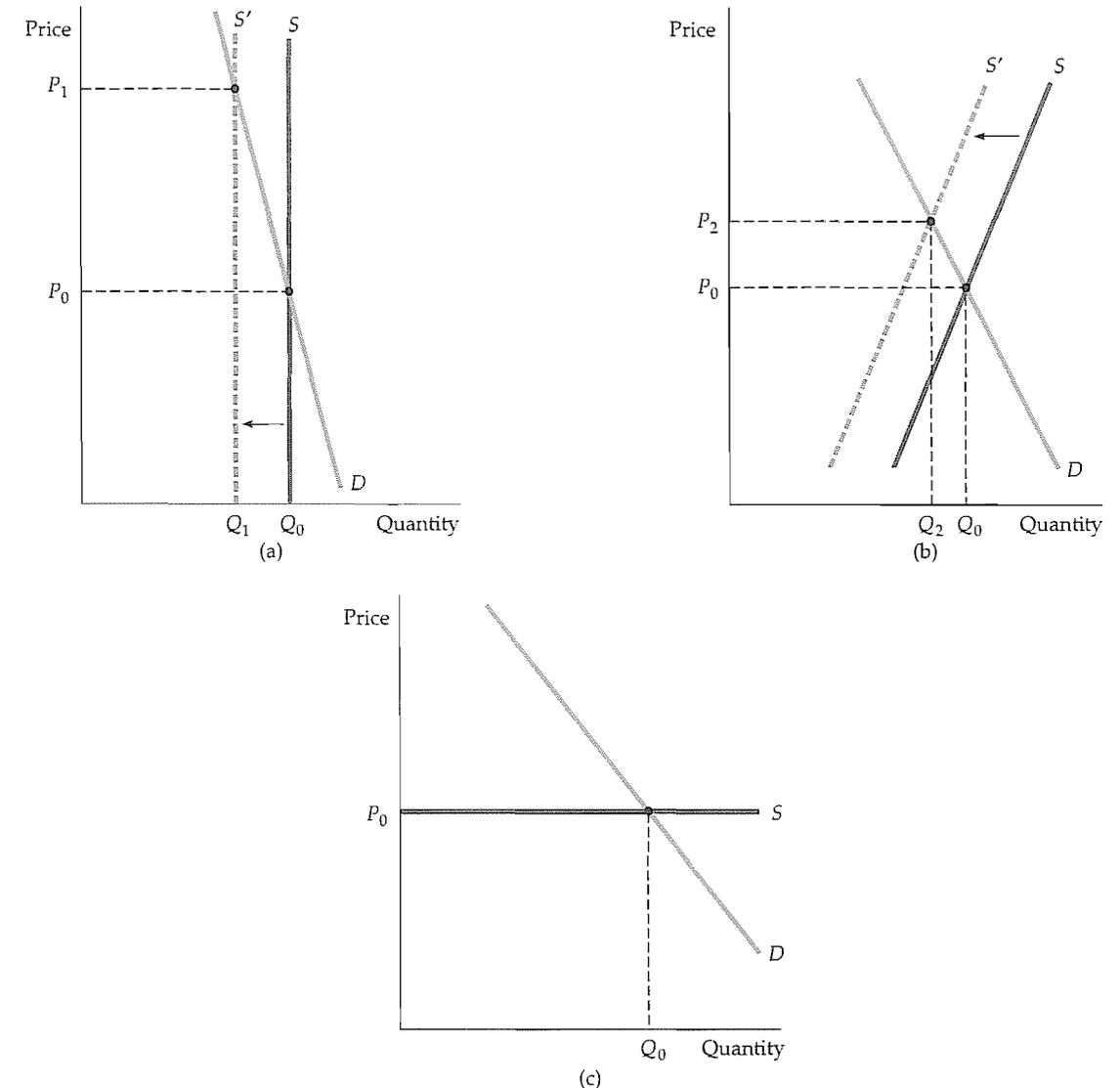


FIGURE 2.17 Supply and Demand for Coffee

(a) A freeze or drought in Brazil causes the supply curve to shift to the left. In the short run, supply is completely inelastic; only a fixed number of coffee beans can be harvested. Demand is also relatively inelastic; consumers change their habits only slowly. As a result, the initial effect of the freeze is a sharp increase in price, from P_0 to P_1 . (b) In the intermediate run, supply and demand are both more elastic; thus price falls part of the way back, to P_2 . (c) In the long run, supply is extremely elastic; because new coffee trees will have had time to mature, the effect of the freeze will have disappeared. Price returns to P_0 .

In the intermediate run—say, one year after the freeze—both supply and demand are more elastic, supply because existing trees can be harvested more intensively (with some decrease in quality), and demand because consumers have had time to change their buying habits. As part (b) shows, although the intermediate-run supply curve also shifts to the left, price has come down from P_1 to P_2 . The quantity supplied has also increased somewhat from the short run, from Q_1 to Q_2 . In the long run shown in part (c), price returns to its normal level because growers have had time to replace trees damaged by the freeze. The long-run supply curve, then, simply reflects the cost of producing coffee, including the costs of land, of planting and caring for the trees, and of a competitive rate of profit.⁹

*2.6 Understanding and Predicting the Effects of Changing Market Conditions

So far our discussion of supply and demand has been largely qualitative. To use supply and demand curves to analyze and predict the effects of changing market conditions, we must begin attaching numbers to them. For example, to see how a 50-percent reduction in the supply of Brazilian coffee may affect the world price of coffee, we must determine actual supply and demand curves and then calculate the shifts in those curves and the resulting changes in price.

In this section, we will see how to do simple “back of the envelope” calculations with linear supply and demand curves. Although they are often approximations of more complex curves, we use linear curves because they are easier to work with. It may come as a surprise, but one can do some informative economic analyses on the back of a small envelope with a pencil and a pocket calculator.

First, we must learn how to “fit” linear demand and supply curves to market data. (By this we do not mean *statistical fitting* in the sense of linear regression or other statistical techniques, which we discuss later in the book.) Suppose we have two sets of numbers for a particular market: The first set consists of the price and quantity that generally prevail in the market (i.e., the price and quantity that prevail “on average,” when the market is in equilibrium, or when market conditions are “normal”). We call these numbers the *equilibrium price* and *quantity* and denote them by P^* and Q^* . The second set consists of the price elasticities of supply and demand for the market (at or near the equilibrium), which we denote by E_S and E_D , as before.

These numbers may come from a statistical study done by someone else; they may be numbers that we simply think are reasonable; or they may be numbers that we want to try out on a “what if” basis. Our goal is to *write down the supply and demand curves that fit (i.e., are consistent with) these numbers*. We can then determine numerically how a change in a variable such as GNP, the price of another good, or some cost of production will cause supply or demand to shift and thereby affect market price and quantity.

⁹ You can learn more about the world coffee market from the Foreign Agriculture Service of the U.S. Department of Agriculture. Their Web site is www.fas.usda.gov/market.html.

Let’s begin with the linear curves shown in Figure 2.18. We can write these curves algebraically as follows:

$$\text{Demand: } Q = a - bP \quad (2.4a)$$

$$\text{Supply: } Q = c + dP \quad (2.4b)$$

Our problem is to choose numbers for the constants a , b , c , and d . This is done, for supply and for demand, in a two-step procedure:

- *Step 1:* Recall that each price elasticity, whether of supply or demand, can be written as

$$E = (P/Q)(\Delta Q/\Delta P),$$

where $\Delta Q/\Delta P$ is the change in quantity demanded or supplied resulting from a small change in price. For linear curves, $\Delta Q/\Delta P$ is constant. From equations (2.4a) and (2.4b), we see that $\Delta Q/\Delta P = d$ for supply and $\Delta Q/\Delta P = -b$ for demand. Now, let’s substitute these values for $\Delta Q/\Delta P$ into the elasticity formula:

$$\text{Demand: } E_D = -b(P^*/Q^*) \quad (2.5a)$$

$$\text{Supply: } E_S = d(P^*/Q^*), \quad (2.5b)$$

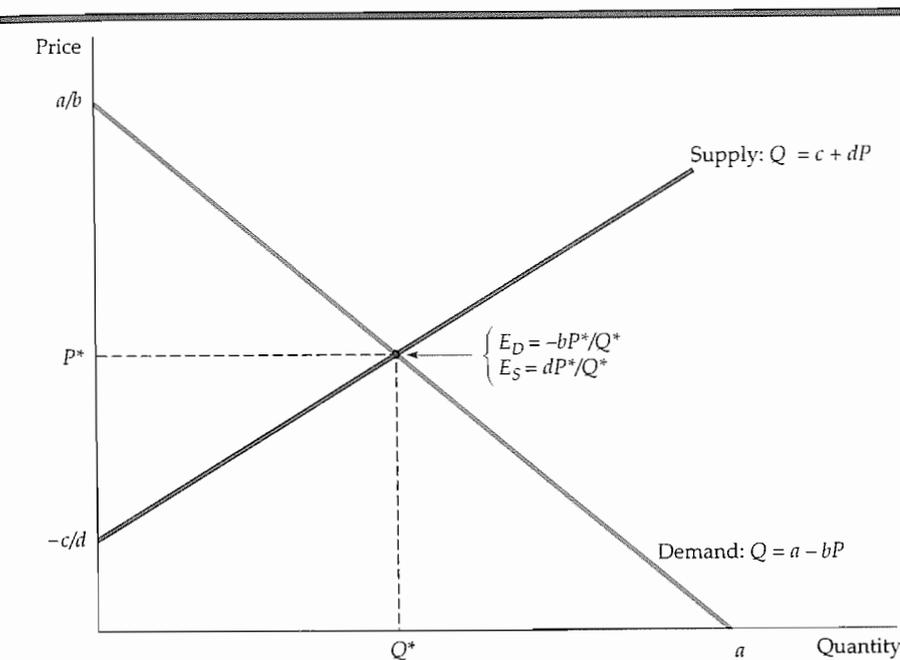


FIGURE 2.18 Fitting Linear Supply and Demand Curves to Data

Linear supply and demand curves provide a convenient tool for analysis. Given data for the equilibrium price and quantity P^* and Q^* , as well as estimates of the elasticities of demand and supply E_D and E_S , we can calculate the parameters c and d for the supply curve and a and b for the demand curve. (In the case drawn here, $c < 0$.) The curves can then be used to analyze the behavior of the market quantitatively.

where P^* and Q^* are the equilibrium price and quantity for which we have data and to which we want to fit the curves. Because we have numbers for E_S , E_D , P^* , and Q^* , we can substitute these numbers in equations (2.5a) and (2.5b) and solve for b and d .

- *Step 2:* Since we now know b and d , we can substitute these numbers, as well as P^* and Q^* , into equations (2.4a) and (2.4b) and solve for the remaining constants a and c . For example, we can rewrite equation (2.4a) as

$$a = Q^* + bP^*$$

and then use our data for Q^* and P^* , together with the number we calculated in Step 1 for b , to obtain a .

Let's apply this procedure to a specific example: long-run supply and demand for the world copper market. The relevant numbers for this market are as follows:¹⁰

Quantity $Q^* = 7.5$ million metric tons per year (mmt/yr)

Price $P^* = 75$ cents per pound

Elasticity of supply $E_S = 1.6$

Elasticity of demand $E_D = -0.8$

(The price of copper has fluctuated during the past decade between 50 cents and more than \$1.30, but 75 cents is a reasonable average price for 1980–1990.)

We begin with the supply curve equation (2.4b) and use our two-step procedure to calculate numbers for c and d . The long-run price elasticity of supply is 1.6, $P^* = .75$, and $Q^* = 7.5$.

- *Step 1:* Substitute these numbers in equation (2.5b) to determine d :

$$1.6 = d(0.75/7.5) = 0.1d,$$

so that $d = 1.6/0.1 = 16$.

- *Step 2:* Substitute this number for d , together with the numbers for P^* and Q^* , into equation (2.4b) to determine c :

$$7.5 = c + (16)(0.75) = c + 12,$$

so that $c = 7.5 - 12 = -4.5$. We now know c and d , so we can write our supply curve:

$$\text{Supply: } Q = -4.5 + 16P$$

We can now follow the same steps for the demand curve equation (2.4a). An estimate for the long-run elasticity of demand is -0.8 . First, substitute this number, as well as the values for P^* and Q^* , into equation (2.5a) to determine b :

$$-0.8 = -b(0.75/7.5) = -0.1b,$$

¹⁰The supply elasticity is for primary supply, as shown in Table 2.3. The demand elasticity is a regionally aggregated number based on Fisher, Cootner, and Baily, "An Econometric Model." Quantities refer to what was then the non-Communist world market.

so that $b = 0.8/0.1 = 8$. Second, substitute this value for b and the values for P^* and Q^* in equation (2.4a) to determine a :

$$7.5 = a - (8)(0.75) = a - 6,$$

so that $a = 7.5 + 6 = 13.5$. Thus, our demand curve is

$$\text{Demand: } Q = 13.5 - 8P$$

To check that we have not made a mistake, let's set the quantity supplied equal to the quantity demanded and calculate the resulting equilibrium price:

$$\text{Supply} = -4.5 + 16P = 13.5 - 8P = \text{Demand}$$

$$16P + 8P = 13.5 + 4.5$$

or $P = 18/24 = 0.75$, which is indeed the equilibrium price with which we began.

Although we have written supply and demand so that they depend only on price, they could easily depend on other variables as well. Demand, for example, might depend on income as well as price. We would then write demand as

$$Q = a - bP + fI, \quad (2.6)$$

where I is an index of aggregate income or GNP. For example, I might equal 1.0 in a base year and then rise or fall to reflect percentage increases or decreases in aggregate income.

For our copper market example, a reasonable estimate for the long-run income elasticity of demand is 1.3. For the linear demand curve (2.6), we can then calculate f by using the formula for the income elasticity of demand: $E = (I/Q)(\Delta Q/\Delta I)$. Taking the base value of I as 1.0, we have

$$1.3 = (1.0/7.5)(f)$$

Thus $f = (1.3)(7.5)/(1.0) = 9.75$. Finally, substituting the values $b = 8$, $f = 9.75$, $P^* = 0.75$, and $Q^* = 7.5$ into equation (2.6), we can calculate that a must equal 3.75.

We have seen how to fit linear supply and demand curves to data. Now, to see how these curves can be used to analyze markets, let's look at Example 2.7, which deals with the behavior of copper prices, and Example 2.8, which concerns the world oil market.

EXAMPLE 2.7 Declining Demand and the Behavior of Copper Prices

After reaching a level of about \$1.00 per pound in 1980, the price of copper fell sharply to about 60 cents per pound in 1986. In real (inflation-adjusted) terms, this price was even lower than during the Great Depression 50 years

earlier. Only in 1988–1989 did prices recover somewhat, largely as a result of strikes by miners in Peru and Canada and disrupted supplies. Figure 2.19 shows the behavior of copper prices in 1965–1999 in both real and nominal terms.

Worldwide recessions in 1980 and 1982 contributed to the decline of copper prices; as mentioned above, the income elasticity of copper demand is about 1.3. But copper demand did not pick up as the industrial economies recovered during the mid-1980s. Instead, the 1980s saw a steep decline in demand.

This decline occurred for two reasons. First, a large part of copper consumption is for the construction of equipment for electric power generation and transmission. But by the late 1970s, the growth rate of electric power generation had fallen dramatically in most industrialized countries. In the United States, for example, the growth rate fell from over 6 percent per annum in the 1960s and early 1970s to less than 2 percent in the late 1970s and 1980s. This decline meant a big drop in what had been a major source of copper demand. Second, in the 1980s, other materials, such as aluminum and fiber optics, were increasingly substituted for copper.

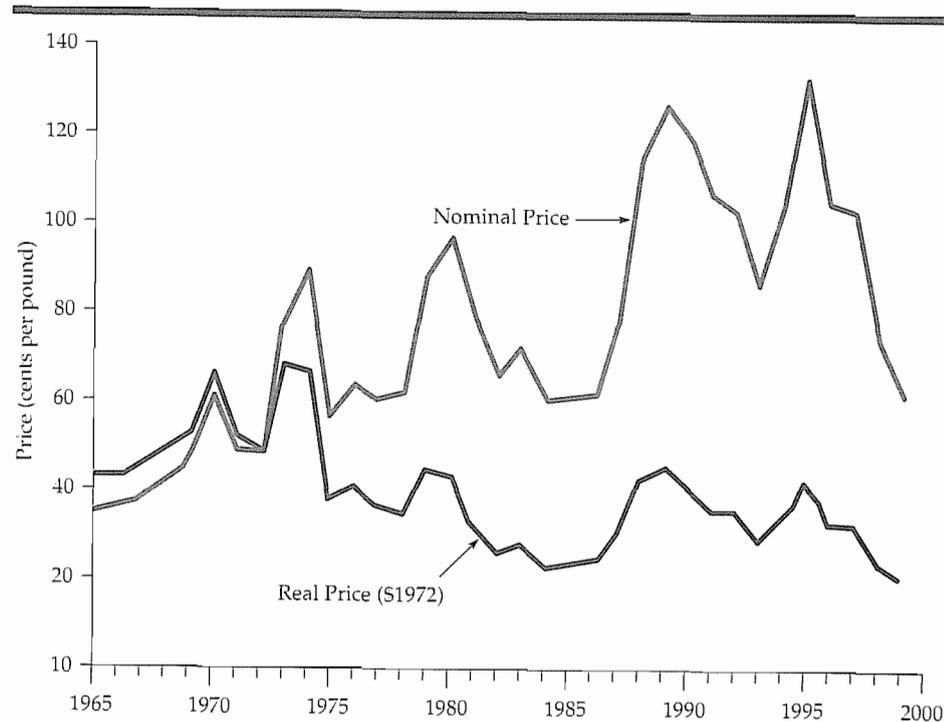


FIGURE 2.19 Copper Prices, 1965–1999

Copper prices are shown in both nominal (no adjustment for inflation) and real (inflation-adjusted) terms. In real terms, copper prices declined steeply from the early 1970s through the mid-1980s as demand fell. In 1988–1990, copper prices rose in response to supply disruptions caused by strikes in Peru and Canada but later fell after the strikes ended. Prices declined sharply during 1996–1999.

Copper producers are concerned about the possible effects of further declines in demand, particularly as strikes end and supplies increase. Declining demand, of course, will depress prices. To find out how much, we can use the linear supply and demand curves that we just derived. Let's calculate the effect on price of a 20-percent decline in demand. Because we are not concerned here with the effects of GNP growth, we can leave the income term I out of demand.

We want to shift the demand curve to the left by 20 percent. In other words, we want the quantity demanded to be 80 percent of what it would be otherwise for every value of price. For our linear demand curve, we simply multiply the right-hand side by 0.8:

$$Q = (0.8)(13.5 - 8P) = 10.8 - 6.4P$$

Supply is again $Q = -4.5 + 16P$. Now we can equate the quantity supplied and the quantity demanded and solve for price:

$$16P + 6.4P = 10.8 + 4.5,$$

or $P = 15.3/22.4 = 68.3$ cents per pound. A decline in demand of 20 percent, therefore, entails a drop in price of roughly 7 cents per pound, or 10 percent.¹¹

EXAMPLE 2.8 Upheaval in the World Oil Market

Since the early 1970s, the world oil market has been buffeted by the OPEC cartel and by political turmoil in the Persian Gulf. In 1974, by collectively restraining output, OPEC (the Organization of Petroleum Exporting Countries) pushed world oil prices well above what they would have been in a competitive market. OPEC could do this because it accounted for much of world oil production. During 1979–1980, oil prices shot up again, as the Iranian revolution and the outbreak of the Iran-Iraq war sharply reduced Iranian and Iraqi production. During the 1980s, the price gradually declined, as demand fell and competitive (i.e., non-OPEC) supply rose in response to price. Prices remained relatively stable during 1988–1999, except for a temporary spike in 1990 following the Iraqi invasion of Kuwait, a decline during 1997–1998, and an increase in 1999. Figure 2.20 shows the world price of oil from 1970 to 1999, in both nominal and real terms.

The Persian Gulf is one of the less stable regions of the world, which has led to concern over the possibility of new oil supply disruptions and sharp increases in oil prices. What would happen to oil prices—in both the short run and longer run—if a war or revolution in the Persian Gulf caused a sharp cut-back in oil production? Let's see how simple supply and demand curves can be used to predict the outcome of such an event.

¹¹ You can obtain recent data and learn more about the behavior of copper prices by accessing the Web site of the U.S. Geological Survey at <http://minerals.usgs.gov/minerals/pubs/commodity/copper>.

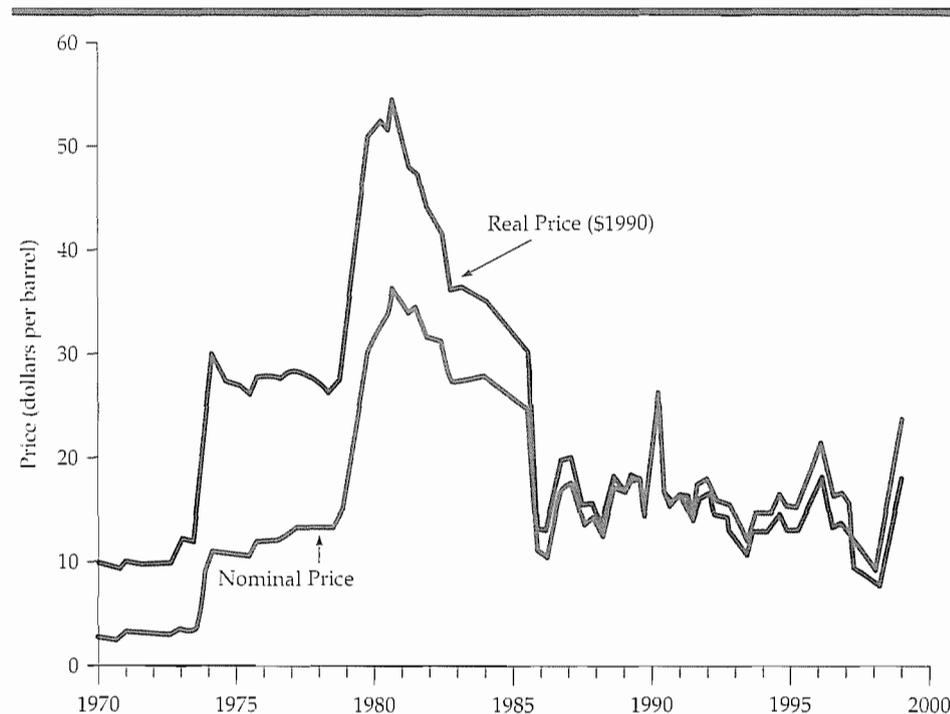


FIGURE 2.20 Price of Crude Oil

The OPEC cartel and political events caused the price of oil to rise sharply at times. It later fell as supply and demand adjusted.

This example is set in 1997, so all prices are measured in 1997 dollars. Here are some rough figures:

- 1997 World price = \$18 per barrel
- World demand and total supply = 23 billion barrels per year (bb/yr)
- 1997 OPEC supply = 10 bb/yr
- Competitive (non-OPEC) supply = 13 bb/yr.¹²

The following table gives price elasticity estimates for oil supply and demand:¹³

	SHORT-RUN	LONG-RUN
World demand:	-0.05	-0.40
Competitive supply:	0.10	0.40

¹²Non-OPEC supply includes the production of China and the former Soviet republics.

¹³For the sources of these numbers and a more detailed discussion of OPEC oil pricing, see Robert S. Pindyck, "Gains to Producers from the Cartelization of Exhaustible Resources," *Review of Economics and Statistics* 60 (May 1978): 238-51; James M. Griffin and David J. Teece, *OPEC Behavior and World Oil Prices* (London: Allen and Unwin, 1982); and Hillard G. Huntington, "Inferred Demand and Supply Elasticities from a Comparison of World Oil Models," in T. Sterner, ed., *International Energy Economics* (London: Chapman and Hall, 1992).

You should verify that these numbers imply the following for demand and competitive supply in the *short run*:

$$\text{Short-run demand: } D = 24.08 - 0.06P$$

$$\text{Short-run competitive supply: } S_C = 11.74 + 0.07P$$

Of course, *total supply* is competitive supply *plus* OPEC supply, which we take as constant at 10 bb/yr. Adding this 10 bb/yr to the competitive supply curve above, we obtain the following for the total short-run supply:

$$\text{Short-run total supply: } S_T = 21.74 + 0.07P$$

You should verify that the quantity demanded and the total quantity supplied are equal at an equilibrium price of \$18 per barrel.

You should also verify that the corresponding demand and supply curves for the *long run* are as follows:

$$\text{Long-run demand: } D = 32.18 - 0.51P$$

$$\text{Long-run competitive supply: } S_C = 7.78 + 0.29P$$

$$\text{Long-run total supply: } S_T = 17.78 + 0.29P$$

Again, you can check that the quantities supplied and demanded equate at a price of \$18.

Saudi Arabia is one of the world's largest oil producers, accounting for roughly 3 bb/yr, which is nearly one third of OPEC production and about 13 percent of total world production. What would happen to the price of oil if, because of war or political upheaval, Saudi Arabia stopped producing oil? We can use our supply and demand curves to find out.

For the *short run*, simply subtract 3 from total supply:

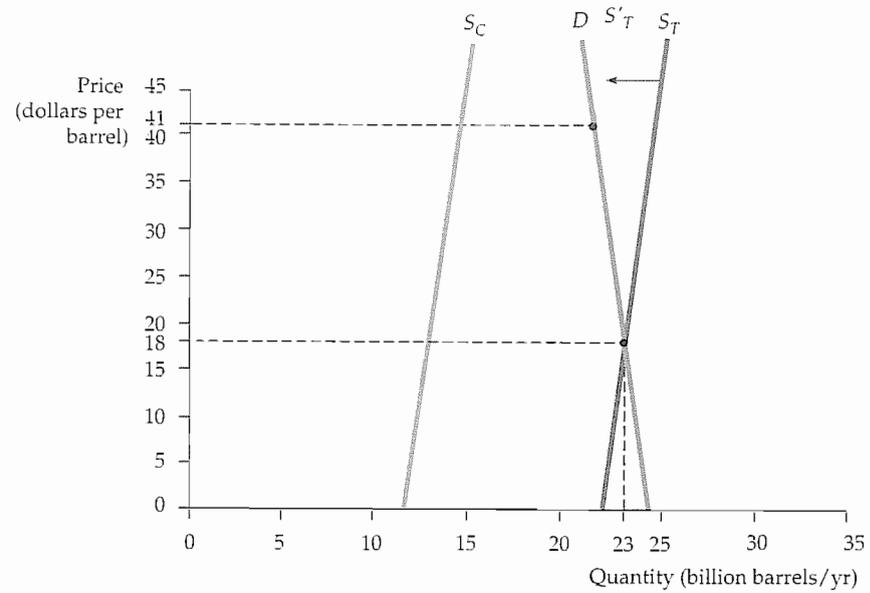
$$\text{Short-run demand: } D = 24.08 - 0.06P$$

$$\text{Short-run total supply: } S_T = 18.74 + 0.07P$$

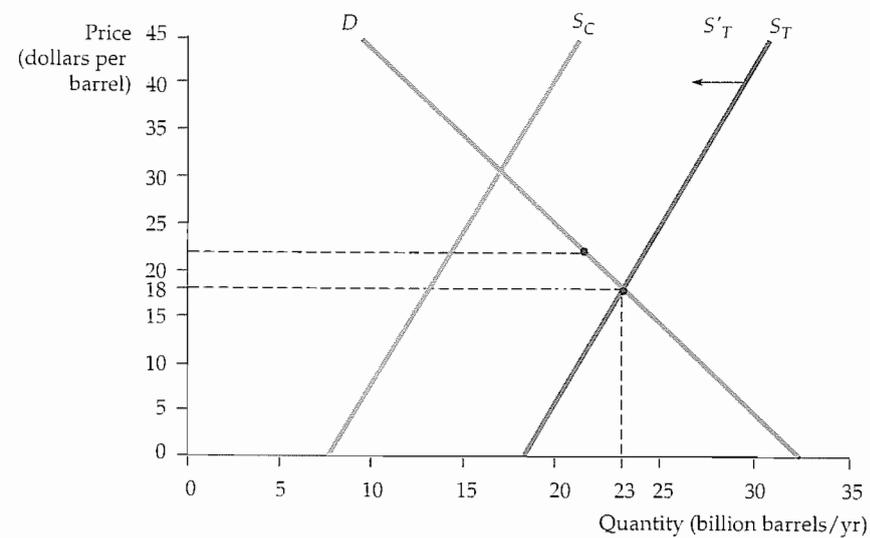
By equating this total quantity supplied with the quantity demanded, we can see that in the short run, the price will more than double to \$41.08 per barrel. Figure 2.21 shows this supply shift and the resulting short-run increase in price. The initial equilibrium is at the intersection of S_T and D . After the drop in Saudi production, the equilibrium occurs where S'_T and D cross.

In the *long run*, however, things will be different. Because both demand and competitive supply are more elastic in the long run, the 3 bb/yr cut in oil production will no longer support such a high price. Subtracting 3 from long-run total supply and equating with long-run demand, we can see that the price will fall to \$21.75, only \$3.75 above the initial \$18 price.

Thus, if Saudi Arabia suddenly stops producing oil, we should expect to see more than a doubling in price. However, we should also expect to see the price gradually decline afterward, as demand falls and competitive supply rises. As



(a)



(b)

FIGURE 2.21 Impact of Saudi Production Cut

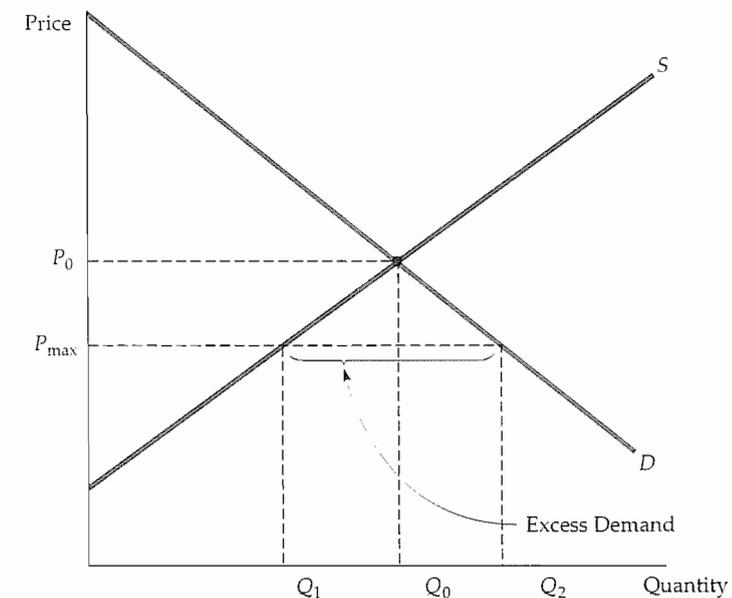
The total supply is the sum of competitive (non-OPEC) supply and the 10 bb/yr of OPEC supply. Part (a) shows the short-run supply and demand curves. If Saudi Arabia stops producing, the supply curve will shift to the left by 3 bb/yr. In the short-run, price will increase sharply. Part (b) shows long-run curves. In the long run, because demand and competitive supply are much more elastic, the impact on price will be much smaller.

Figure 2.20 shows, this is indeed what happened following the sharp decline in Iranian and Iraqi production in 1979–1980. History may or may not repeat itself, but if it does, we can at least predict the impact on oil prices.¹⁴

2.7 Effects of Government Intervention—Price Controls

In the United States and most other industrial countries, markets are rarely free of government intervention. Besides imposing taxes and granting subsidies, governments often regulate markets (even competitive markets) in a variety of ways. In this section we will see how to use supply and demand curves to analyze the effects of one common form of government intervention: price controls. Later, in Chapter 9, we will examine the effects of price controls and other forms of government intervention and regulation in more detail.

Figure 2.22 illustrates the effects of price controls. Here, P_0 and Q_0 are the equilibrium price and quantity that would prevail without government regulation.

**FIGURE 2.22 Effects of Price Controls**

Without price controls, the market clears at the equilibrium price and quantity P_0 and Q_0 . If price is regulated to be no higher than P_{\max} , the quantity supplied falls to Q_1 , the quantity demanded increases to Q_2 , and a shortage develops.

¹⁴ You can obtain recent data and learn more about the world oil market by accessing the Web sites of the American Petroleum Institute at www.api.org or the U.S. Energy Information Administration at www.eia.doe.gov.

The government, however, has decided that P_0 is too high and mandated that the price can be no higher than a maximum allowable *ceiling price*, denoted by P_{\max} . What is the result? At this lower price, producers (particularly those with higher costs) will produce less, and the quantity supplied will drop to Q_1 . Consumers, on the other hand, will demand more at this low price; they would like to purchase the quantity Q_2 . Demand therefore exceeds supply, and a shortage develops—i.e., there is *excess demand*. The amount of excess demand is $Q_2 - Q_1$.

This excess demand sometimes takes the form of queues, as when drivers lined up to buy gasoline during the winter of 1974 and the summer of 1979. In both instances, the lines were the result of price controls; the government prevented domestic oil and gasoline prices from rising along with world oil prices. Sometimes excess demand results in curtailments and supply rationing, as with natural gas price controls and the resulting gas shortages of the mid-1970s, when industrial consumers closed factories because gas supplies were cut off. Sometimes it spills over into other markets, where it artificially increases demand. For example, natural gas price controls caused potential buyers of gas to use oil instead.

Some people gain and some lose from price controls. As Figure 2.22 suggests, producers lose: They receive lower prices, and some leave the industry. Some but not all consumers gain. While those who can purchase the good at a lower price are better off, those who have been “rationed out” and cannot buy the good at all are worse off. How large are the gains to the winners and how large are the losses to the losers? Do total gains exceed total losses? To answer these questions we need a method to measure the gains and losses from price controls and other forms of government intervention. We discuss such a method in Chapter 9.

EXAMPLE 2.9 Price Controls and Natural Gas Shortages

In 1954, the federal government began regulating the wellhead price of natural gas. Initially the controls were not binding; the ceiling prices were above those that cleared the market. But in about 1962, when these ceiling prices did become binding, excess demand for natural gas developed and slowly began to grow. In the 1970s, this excess demand, spurred by higher oil prices, became severe and led to widespread curtailments. Soon ceiling prices were far below prices that would have prevailed in a free market.¹⁵

Today, producers and industrial consumers of natural gas, oil, and other commodities are concerned that the government might respond, once again, with price controls if prices rise sharply. To understand the likely impact of such price controls, we will go back to the year 1975 and calculate the impact of natural gas price controls at that time.

¹⁵This regulation began with the Supreme Court’s 1954 decision requiring the then Federal Power Commission to regulate wellhead prices on natural gas sold to interstate pipeline companies. These price controls were largely removed during the 1980s, under the mandate of the Natural Gas Policy Act of 1978. For a detailed discussion of natural gas regulation and its effects, see Paul W. MacAvoy and Robert S. Pindyck, *The Economics of the Natural Gas Shortage* (Amsterdam: North-Holland, 1975); R. S. Pindyck, “Higher Energy Prices and the Supply of Natural Gas,” *Energy Systems and Policy* 2 (1978): 177–209; and Arlon R. Tussing and Connie C. Barlow, *The Natural Gas Industry* (Cambridge, MA: Ballinger, 1984).

Based on econometric studies of natural gas markets and the behavior of those markets as controls were gradually lifted during the 1980s, the following data describe the market in 1975.

- The free-market price of natural gas would have been about \$2.00 per mcf (thousand cubic feet);
- Production and consumption would have been about 20 Tcf (trillion cubic feet);
- The average price of oil (including both imports and domestic production), which affects both supply and demand for natural gas, was about \$8/barrel.

A reasonable estimate for the price elasticity of supply is 0.2. Higher oil prices also lead to more natural gas production because oil and gas are often discovered and produced together; an estimate of the cross-price elasticity of supply is 0.1. As for demand, the price elasticity is about -0.5 , and the cross-price elasticity with respect to oil price is about 1.5. You can verify that the following linear supply and demand curves fit these numbers:

$$\text{Supply: } Q = 14 + 2P_G + .25P_O$$

$$\text{Demand: } Q = -5P_G + 3.75P_O,$$

where Q is the quantity of natural gas (in Tcf), P_G is the price of natural gas (in dollars per mcf), and P_O is the price of oil (in dollars per barrel). You can also verify, by equating the quantities supplied and demanded and substituting \$8.00 for P_O , that these supply and demand curves imply an equilibrium free market price of \$2.00 for natural gas.

The regulated price of gas in 1975 was about \$1.00 per mcf. Substituting this price for P_G in the supply function gives a quantity supplied (Q_1 in Figure 2.22) of 18 Tcf. Substituting for P_G in the demand function gives a demand (Q_2 in Figure 2.22) of 25 Tcf. Price controls thus created an excess demand of $25 - 18 = 7$ Tcf, which manifested itself in the form of widespread curtailments.

Price regulation was a major component of U.S. energy policy during the 1960s and 1970s, and continued to influence the evolution of natural gas markets in the 1980s. In Example 9.1 of Chapter 9, we show how to measure the gains and losses that result from price controls.

SUMMARY

1. Supply-demand analysis is a basic tool of microeconomics. In competitive markets, supply and demand curves tell us how much will be produced by firms and how much will be demanded by consumers as a function of price.
2. The market mechanism is the tendency for supply and demand to equilibrate (i.e., for price to move to

the market-clearing level), so that there is neither excess demand nor excess supply.

3. Elasticities describe the responsiveness of supply and demand to changes in price, income, or other variables. For example, the price elasticity of demand measures the percentage change in the quantity demanded resulting from a 1-percent increase in price.

- Elasticities pertain to a time frame, and for most goods it is important to distinguish between short-run and long-run elasticities.
- If we can estimate, at least roughly, the supply and demand curves for a particular market, we can calculate the market-clearing price by equating the quantity supplied with the quantity demanded. Also, if we know how supply and demand depend on other economic variables, such as income or the prices of other goods, we can calculate how the market-clearing

price and quantity will change as these other variables change. This is a means of explaining or predicting market behavior.

- Simple numerical analyses can often be done by fitting linear supply and demand curves to data on price and quantity and to estimates of elasticities. For many markets, such data and estimates are available, and simple "back of the envelope" calculations can help us understand the characteristics and behavior of the market.

QUESTIONS FOR REVIEW

- Suppose that unusually hot weather causes the demand curve for ice cream to shift to the right. Why will the price of ice cream rise to a new market-clearing level?
- Use supply and demand curves to illustrate how each of the following events would affect the price of butter and the quantity of butter bought and sold: (a) an increase in the price of margarine; (b) an increase in the price of milk; (c) a decrease in average income levels.
- Suppose a 3-percent increase in the price of corn flakes causes a 6-percent decline in the quantity demanded. What is the elasticity of demand for corn flakes?
- Why do long-run elasticities of demand differ from short-run elasticities? Consider two goods: paper towels and televisions. Which is a durable good? Would you expect the price elasticity of demand for paper towels to be larger in the short run or in the long run? Why? What about the price elasticity of demand for televisions?
- Explain why for many goods, the long-run price elasticity of supply is larger than the short-run elasticity.
- Suppose the government regulates the prices of beef and chicken and sets them below their market-clearing levels. Explain why shortages of these goods will develop and what factors will determine the sizes of the shortages. What will happen to the price of pork? Explain briefly.
- In a discussion of tuition rates, a university official argues that the demand for admission is completely price inelastic. As evidence, she notes that while the university has doubled its tuition (in real terms) over the past 15 years, neither the number nor quality of students applying has decreased. Would you accept this argument? Explain briefly. (*Hint:* The official makes an assertion about the demand for admission, but does she actually observe a demand curve? What else could be going on?)
- Use supply and demand curve shifts to illustrate the effect of the following events on the market for apples. Make clear the direction of the change in both price and quantity sold.
 - Scientists find that an apple a day does indeed keep the doctor away.
 - The price of oranges triples.
 - A drought shrinks the apple crop to one-third its normal size.
 - Thousands of college students abandon the academic life to become apple pickers.
 - Thousands of college students abandon the academic life to become apple growers.
- Suppose the demand curve for a product is given by

$$Q = 10 - 2P + P_S$$
 where P is the price of the product and P_S is the price of a substitute good. The price of the substitute good is \$2.00.
 - Suppose $P = \$1.00$. What is the price elasticity of demand? What is the cross-price elasticity of demand?
 - Suppose the price of the good, P , goes to \$2.00. Now what is the price elasticity of demand, and what is the cross-price elasticity of demand?
- Suppose that rather than the declining demand assumed in Example 2.7, a decrease in the cost of copper production causes the supply curve to shift to the right by 40 percent. How will the price of copper change?
- Suppose the demand for natural gas is perfectly inelastic. What would be the effect, if any, of natural gas price controls?

EXERCISES

- Consider a competitive market for which the quantities demanded and supplied (per year) at various prices are given as follows:

PRICE (DOLLARS)	DEMAND (MILLIONS)	SUPPLY (MILLIONS)
60	22	14
80	20	16
100	18	18
120	16	20

- Calculate the price elasticity of demand when the price is \$80, and when the price is \$100.
 - Calculate the price elasticity of supply when the price is \$80 and when the price is \$100.
 - What are the equilibrium price and quantity?
 - Suppose the government sets a price ceiling of \$80. Will there be a shortage, and if so, how large will it be?
- Refer to Example 2.4 on the market for wheat. At the end of 1998, both Brazil and Indonesia opened their wheat markets to U.S. farmers. (Source: <http://www.fas.usda.gov/>) Suppose that these new markets add 200 million bushels to U.S. wheat demand. What will be the free-market price of wheat and what quantity will be produced and sold by U.S. farmers in this case?
 - A vegetable fiber is traded in a competitive world market, and the world price is \$9 per pound. Unlimited quantities are available for import into the United States at this price. The U.S. domestic supply and demand for various price levels are shown below.

PRICE	U.S. SUPPLY (MILLION LBS)	U.S. DEMAND (MILLION LBS)
3	2	34
6	4	28
9	6	22
12	8	16
15	10	10
18	12	4

- What is the equation for demand? What is the equation for supply?
- At a price of \$9, what is the price elasticity of demand? What is it at a price of \$12?
- What is the price elasticity of supply at \$9? At \$12?

- In a free market, what will be the U.S. price and level of fiber imports?
- The rent control agency of New York City has found that aggregate demand is $Q_D = 100 - 5P$. Quantity is measured in tens of thousands of apartments. Price, the average monthly rental rate, is measured in hundreds of dollars. The agency also noted that the increase in Q at lower P results from more three-person families coming into the city from Long Island and demanding apartments. The city's board of realtors acknowledges that this is a good demand estimate and has shown that supply is $Q_S = 50 + 5P$.
 - If both the agency and the board are right about demand and supply, what is the free-market price? What is the change in city population if the agency sets a maximum average monthly rent of \$100 and all those who cannot find an apartment leave the city?
 - Suppose the agency bows to the wishes of the board and sets a rental of \$900 per month on all apartments to allow landlords a "fair" rate of return. If 50 percent of any long-run increases in apartment offerings comes from new construction, how many apartments are constructed?
 - Much of the demand for U.S. agricultural output has come from other countries. From Example 2.4, total demand is $Q = 3244 - 283P$. In addition, we are told that domestic demand is $Q_D = 1700 - 107P$. Domestic supply is $Q_S = 1944 + 207P$. Suppose the export demand for wheat falls by 40 percent.
 - U.S. farmers are concerned about this drop in export demand. What happens to the free-market price of wheat in the United States? Do the farmers have much reason to worry?
 - Now, suppose the U.S. government wants to buy enough wheat to raise the price to \$3.50 per bushel. With this drop in export demand, how much wheat would the government have to buy? How much would this cost the government?
 - In 1998, Americans smoked 470 billion cigarettes. The average retail price was \$2 per pack. Statistical studies have shown that the price elasticity of demand is -0.4 , and the price elasticity of supply is 0.5 . Using this information, derive linear demand and supply curves for the cigarette market. (For more information on this market, see Frank J. Chaloupka, "The Economics of Smoking," NBER working paper, 1999, which can be accessed on the Web at <http://nberws.nber.org/papers/W7047.pdf>).
 - In Example 2.7 we examined the effect of a 20-percent decline in copper demand on the price of copper, using the linear supply and demand curves developed in Section 2.4. Suppose the long-run price elasticity of copper demand were -0.4 instead of -0.8 .

- a. Assuming, as before, that the equilibrium price and quantity are $P^* = 75$ cents per pound and $Q^* = 7.5$ million metric tons per year, derive the linear demand curve consistent with the smaller elasticity.
- b. Using this demand curve, recalculate the effect of a 20-percent decline in copper demand on the price of copper.
8. Example 2.8 analyzes the world oil market. Using the data given in that example:
- a. Show that the short-run demand and competitive supply curves are indeed given by

$$D = 24.08 - 0.06P$$

$$S_C = 11.74 + 0.07P$$

- b. Show that the long-run demand and competitive supply curves are indeed given by

$$D = 32.18 - 0.51P$$

$$S_C = 7.78 + 0.29P$$

- c. During the late 1990s, Saudi Arabia accounted for 3 billion barrels per year of OPEC's production. Suppose that war or revolution caused Saudi Arabia to stop producing oil. Use the model above to calculate what would happen to the price of oil in the short run *and* the long run if OPEC's production were to drop by 3 billion barrels per year.

9. Refer to Example 2.9, which analyzes the effects of price controls on natural gas.
- a. Using the data in the example, show that the following supply and demand curves did indeed describe the market in 1975:

$$\text{Supply: } Q = 14 + 2P_G + 0.25P_O$$

$$\text{Demand: } Q = -5P_G + 3.75P_O,$$

where P_G and P_O are the prices of natural gas and oil, respectively. Also verify that if the price of oil is \$8.00, these curves imply a free-market price of \$2.00 for natural gas.

- b. Suppose the regulated price of gas in 1975 had been \$1.50 per thousand cubic feet instead of \$1.00. How much excess demand would there have been?
- c. Suppose that the market for natural gas had *not* been regulated. If the price of oil had increased from \$8.00 to \$16.00, what would have happened to the free market price of natural gas?
- *10. The table below shows the retail price and sales for instant coffee and roasted coffee for 1997 and 1998.
- a. Using this data alone, estimate the short-run price elasticity of demand for roasted coffee. Also, derive a linear demand curve for roasted coffee.
- b. Now estimate the short-run price elasticity of demand for instant coffee. Derive a linear demand curve for instant coffee.
- c. Which coffee has the higher short-run price elasticity of demand? Why do you think this is the case?

YEAR	RETAIL PRICE OF INSTANT COFFEE (\$/LB)	SALES OF INSTANT COFFEE (MILLION LBS)	RETAIL PRICE OF ROASTED COFFEE (\$/LB)	SALES OF ROASTED COFFEE (MILLION LBS)
1997	10.35	75	4.11	820
1998	10.48	70	3.76	850

PART 2

Producers, Consumers, and Competitive Markets

CHAPTERS

- 3 Consumer Behavior 61
- 4 Individual and Market Demand 101
- 5 Choice under Uncertainty 149
- 6 Production 177
- 7 The Cost of Production 203
- 8 Profit Maximization and Competitive Supply 251
- 9 The Analysis of Competitive Markets 287

PART 2 presents the theoretical core of microeconomics.

Chapters 3 and 4 explain the principles underlying consumer demand. We see how consumers make consumption decisions, how their preferences and budget constraints determine their demands for various goods, and why different goods have different demand characteristics. Chapter 5 contains more advanced material that shows how to analyze consumer choice under uncertainty. We explain why people usually dislike risky situations, and show how they can reduce risk, and how they choose among risky alternatives.

Chapters 6 and 7 develop the theory of the firm. We see how firms combine inputs, such as capital, labor, and raw materials, to produce goods and services in a way that minimizes the costs of production. We also see how a firm's costs depend on its rate of production and production experience. Chapter 8 then shows how firms choose profit-maximizing rates of production. We also see how the production decisions of individual firms combine to determine the competitive market supply curve and its characteristics.

Chapter 9 applies supply and demand curves to the analysis of competitive markets. We show how government policies, such as price controls, quotas, taxes, and subsidies, can have wide-ranging effects on consumers and producers and explain how supply-demand analysis can be used to evaluate these effects.

CHAPTER 3

Consumer Behavior

A few years ago, General Mills decided to introduce a new product. The new brand, Apple-Cinnamon Cheerios, offered a sweetened and more flavorful variant on General Mills' classic Cheerios product. But before Apple-Cinnamon Cheerios could be extensively marketed, the company had to resolve an important problem: *How high a price should it charge?* No matter how good the cereal was, its profitability would be affected considerably by the company's pricing decision. Knowing that consumers would pay more for a new product with added ingredients was not enough. The question was *how much more*. General Mills, therefore, had to conduct a careful analysis of consumer preferences to determine the demand for Apple-Cinnamon Cheerios.

General Mills' problem in determining consumer preferences mirrors the more complex problem faced by the U.S. Congress in evaluating the federal Food Stamps program. The goal of the program is to give to low-income households coupons that can be exchanged for food. But there has always been a problem in the program's design that complicates its assessment: To what extent do food stamps provide people with *more* food, as opposed to simply subsidizing the purchase of food that they would have bought anyway? In other words, has the program turned out to be little more than an income supplement that people spend largely on nonfood items instead of a solution to the nutritional problems of the poor? As in the cereal example, an analysis of consumer behavior is needed. In this case, the federal government must determine how spending on food, as opposed to spending on other goods, is affected by changing income levels and prices.

Solving these two problems—one involving corporate policy and the other public policy—requires an understanding of the **theory of consumer behavior**: the explanation of how consumers allocate incomes to the purchase of different goods and services.

Consumer Behavior

How can a consumer with a limited income decide which goods and services to buy? This is a fundamental issue in microeconomics—one that we address in this chapter and the next. We will see how consumers allocate their incomes across goods and explain how these allocation decisions determine

Chapter Outline

- 3.1 Consumer Preferences 62
- 3.2 Budget Constraints 75
- 3.3 Consumer Choice 79
- 3.4 Revealed Preference 86
- 3.5 Marginal Utility and Consumer Choice 89
- *3.6 Cost-of-Living Indexes 92

List of Examples

- 3.1 Designing New Automobiles (I) 71
- 3.2 Designing New Automobiles (II) 81
- 3.3 Decision Making and Public Policy 82
- 3.4 A College Trust Fund 85
- 3.5 Revealed Preference for Recreation 88
- 3.6 Gasoline Rationing 91
- 3.7 The Bias in the CPI 97

theory of consumer behavior
Description of how consumers allocate incomes among different goods and services to maximize their well-being.

the demands for various goods and services. In turn, understanding consumer purchasing decisions will help us to understand how changes in income and prices affect demands for goods and services and why the demands for some products are more sensitive than others to changes in prices and income.

Consumer behavior is best understood in three distinct steps:

1. **Consumer Preferences:** The first step is to find a practical way to describe the reasons people might prefer one good to another. We will see how a consumer's preferences for various goods can be described graphically and algebraically.
2. **Budget Constraints:** Of course, consumers also consider *prices*. In Step 2, therefore, we take into account the fact that consumers have limited incomes which restrict the quantities of goods they can buy. What does a consumer do in this situation? We find the answer to this question by putting consumer preferences and budget constraints together in the third step.
3. **Consumer Choices:** Given their preferences and limited incomes, consumers choose to buy combinations of goods that maximize their satisfaction. These combinations will depend on the prices of various goods. Thus understanding consumer choice will help us understand *demand*—i.e., how the quantity of a good that consumers choose to purchase depends on its price.

These three steps are the basics of consumer theory, and we will go through them in detail in the first three sections of this chapter. Afterward, we will explore a number of other interesting aspects of consumer behavior. For example, we will see how one can determine the nature of consumer preferences from actual observations of consumer behavior. Thus if a consumer chooses one good over a similarly priced alternative, we can infer that he or she prefers the first good. Similar kinds of conclusions can be drawn from the actual decisions that consumers make in response to changes in the prices of the various goods and services that are available for purchase.

At the end of this chapter, we will return to the discussion of real and nominal prices that we began in Chapter 1. We saw that the Consumer Price Index can provide one measure of how the well-being of consumers changes over time. In this chapter, we delve more deeply into the subject of purchasing power by describing a range of indexes that measure changes in purchasing power over time. Because they affect the benefits and costs of numerous social-welfare programs, these indexes are significant tools in setting government policy in the United States.

3.1 Consumer Preferences

Given both the vast number of goods and services that our industrial economy provides for purchase and the diversity of personal tastes, how can we describe consumer preferences in a coherent way? Let's begin by thinking about how a consumer might compare different groups of items available for purchase. Will one group of items be preferred to another group? Or will the consumer be indifferent between the two groups?

Market Baskets

We use the term *market basket* to refer to such a group of items. Specifically, a **market basket** is a list with specific quantities of one or more commodities. A market basket might contain the various food items in a grocery cart. It might

market basket (or bundle)
List with specific quantities of one or more goods.

TABLE 3.1 Alternative Market Baskets

MARKET BASKET	UNITS OF FOOD	UNITS OF CLOTHING
A	20	30
B	10	50
D	40	20
E	30	40
G	10	20
H	10	40

Note: We will avoid the use of the letters C and F to represent market baskets, whenever market baskets might be confused with the number of units of food and clothing.

also refer to the quantities of food, clothing, and housing that a consumer buys each month. Many economists also use the word **bundle** to mean the same thing as market basket.

How do consumers select market baskets? How do they decide, for example, how much food versus clothing to buy each month? Although selections may occasionally be arbitrary, as we will soon see, consumers usually select market baskets that make them as well off as possible.

Table 3.1 shows several market baskets consisting of various amounts of food and clothing purchased on a monthly basis. The number of food items can be measured in any number of ways: by total number of containers, by number of packages of each item (e.g., milk, meat, etc.), or by number of pounds or grams. Likewise, clothing can be counted as total number of pieces, as number of pieces of each type of clothing, or as total weight or volume. Because the method of measurement is largely arbitrary, we will simply describe the items in a market basket in terms of the total number of *units* of each commodity. Market basket A, for example, consists of 20 units of food and 30 units of clothing, basket B of 10 units of food and 50 units of clothing, and so on.

To explain the theory of consumer behavior, we will ask whether consumers *prefer* one market basket to another. Note that the theory assumes that consumers' preferences are consistent and make sense. We explain what we mean by these assumptions in the next subsection.

Some Basic Assumptions about Preferences

The theory of consumer behavior begins with three basic assumptions about people's preferences for one market basket versus another. We believe that these assumptions hold for most people in most situations:

1. **Completeness:** Preferences are assumed to be *complete*. In other words, consumers can compare and rank all possible baskets. Thus, for any two market baskets A and B, a consumer will prefer A to B, will prefer B to A, or will be indifferent between the two. By *indifferent* we mean that a person will be equally satisfied with either basket. Note that these preferences ignore costs. A consumer might prefer steak to hamburger but buy hamburger because it is cheaper.
2. **Transitivity:** Preferences are *transitive*. Transitivity means that if a consumer prefers basket A to basket B and basket B to basket C, then the consumer also prefers A to C. For example, if a Porsche is preferred to a

Cadillac and a Cadillac to a Chevrolet, then a Porsche is also preferred to a Chevrolet. Transitivity is normally regarded as necessary for consumer consistency.

3. **More is better than less:** Goods are assumed to be desirable—i.e., to be *good*. Consequently, *consumers always prefer more of any good to less*. In addition, consumers are never satisfied or satiated; *more is always better, even if just a little better*.¹ This assumption is made for pedagogic reasons; namely, it simplifies the graphical analysis. Of course, some goods, such as air pollution, may be undesirable, and consumers will always prefer less. We ignore these “bads” in the context of our immediate discussion of consumer choice because most consumers would not choose to purchase them. We will, however, discuss them later in the chapter.

These three assumptions form the basis of consumer theory. They do not explain consumer preferences, but they do impose a degree of rationality and reasonableness on them. Building on these assumptions, we will now explore consumer behavior in greater detail.

Indifference Curves

We can show a consumer's preferences graphically with the use of *indifference curves*. An **indifference curve** represents all combinations of market baskets that provide a person with the same level of satisfaction. That person is therefore *indifferent* among the market baskets represented by the points graphed on the curve.

indifference curve Curve representing all combinations of market baskets that provide a consumer with the same level of satisfaction.

Given our three assumptions about preferences, we know that a consumer can always indicate either a preference for one market basket over another or indifference between the two. We can then use this information to rank all possible consumption choices. In order to appreciate this principle in graphic form, let's assume that there are only two goods available for consumption: food F and clothing C . In this case, all market baskets describe combinations of food and clothing that a person might wish to consume. As we have already seen, Table 3.1 provides some examples of baskets containing various amounts of food and clothing.

In order to graph a consumer's indifference curve, it helps first to graph his or her individual preferences. Figure 3.1 shows the same baskets listed in Table 3.1. The horizontal axis measures the number of units of food purchased each week; the vertical axis measures the number of units of clothing. Market basket A , with 20 units of food and 30 units of clothing, is preferred to basket G because A contains more food *and* more clothing (recall our third assumption that more is better than less). Similarly, market basket E , which contains even more food and even more clothing, is preferred to A . In fact, we can easily compare all market baskets in the two shaded areas (such as E and G) to A because they all contain either more or less of both food and clothing. Note, however, that B contains more clothing but less food than A . Likewise, D contains more food but less clothing than A . Therefore, comparisons of market basket A with baskets B , D , and H are not possible without more information about the consumer's ranking.

This additional information is provided in Figure 3.2, which shows an indifference curve, labeled U_1 , that passes through points A , B , and D . This curve indicates that the consumer is indifferent among these three market baskets. It

¹ Thus some economists use the term *nonsatiation* to refer to this third assumption.

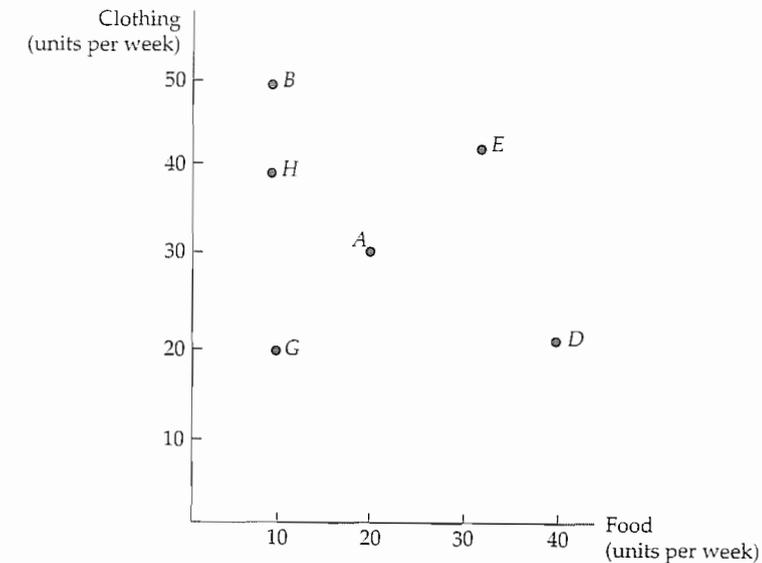


FIGURE 3.1 Describing Individual Preferences

Because more of each good is preferred to less, we can compare market baskets in the shaded areas. Basket A is clearly preferred to basket G , while E is clearly preferred to A . However, A cannot be compared with B , D , or H without additional information.

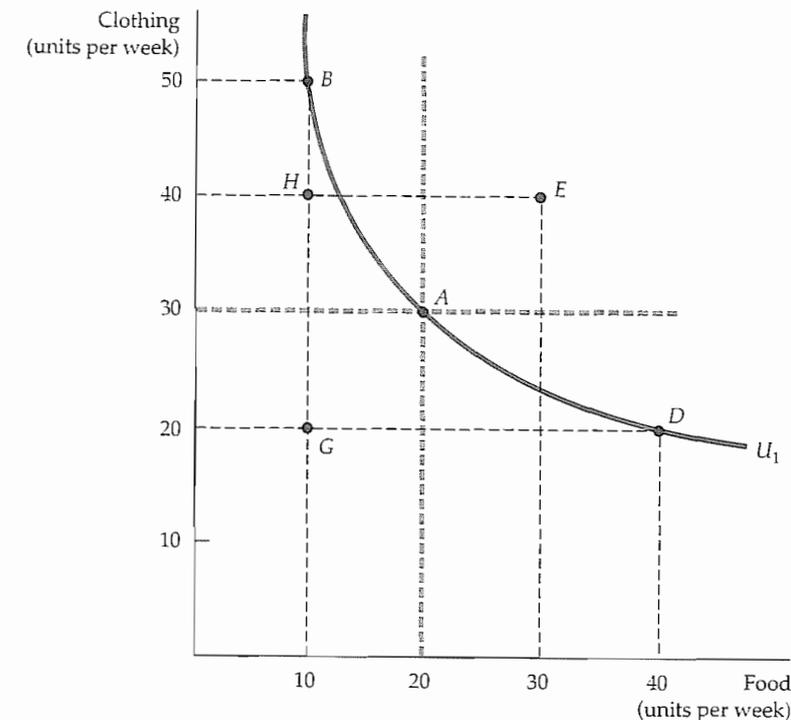


FIGURE 3.2 An Indifference Curve

The indifference curve U_1 that passes through market basket A shows all baskets that give the consumer the same level of satisfaction as does market basket A ; these include baskets B and D . Our consumer prefers basket E , which lies above U_1 , to A , but prefers A to H or G , which lie below U_1 .

tells us that in moving from market basket A to market basket B , the consumer feels neither better nor worse off in giving up 10 units of food to obtain 20 additional units of clothing. Likewise, the consumer is indifferent between points A and D : He or she will give up 10 units of clothing to obtain 20 units of food. On the other hand, the consumer prefers A to H , which lies below U_1 .

Note that the indifference curve in Figure 3.2 slopes downward from left to right. To understand why this must be the case, suppose instead that it sloped upward from A to E . This would violate the assumption that more of any commodity is preferred to less. Because market basket E has more of both food and clothing than market basket A , it must be preferred to A and therefore cannot be on the same indifference curve as A . In fact, any market basket lying *above and to the right* of indifference curve U_1 in Figure 3.2 is preferred to any market basket on U_1 .

Indifference Maps

To describe a person's preferences for *all* combinations of food and clothing, we can graph a set of indifference curves called an **indifference map**. Each indifference curve in the map shows the market baskets among which the person is indifferent. Figure 3.3 shows three indifference curves that form part of an indifference map. Indifference curve U_3 generates the highest level of satisfaction, followed by indifference curves U_2 and U_1 .

Indifference curves cannot intersect. To see why, we will assume the contrary and see how the resulting graph violates our assumptions about consumer behavior. Figure 3.4 shows two indifference curves, U_1 and U_2 , that intersect at A . Because A and B are both on indifference curve U_1 , the consumer must be indifferent between these two market baskets. Because both A and D lie on indifference curve U_2 , the consumer must be indifferent between these market baskets. Consequently, the consumer must also be indifferent between B and D . But this conclusion can't be true: Market basket B must be preferred to D because it

indifference map Graph containing a set of indifference curves showing the market baskets among which a consumer is indifferent.

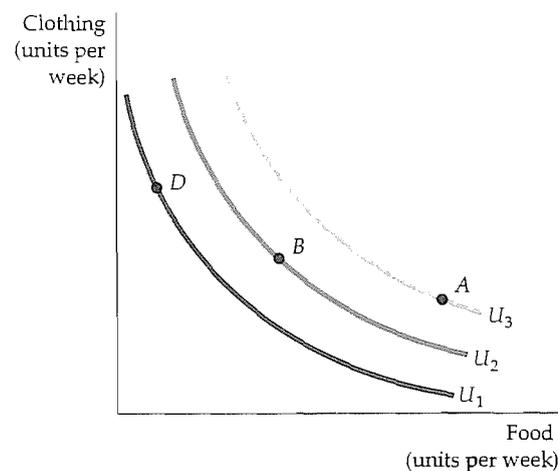


FIGURE 3.3 An Indifference Map

An indifference map is a set of indifference curves that describes a person's preferences. Any market basket on indifference curve U_3 , such as basket A , is preferred to any basket on curve U_2 (e.g., basket B), which in turn is preferred to any basket on U_1 , such as D .

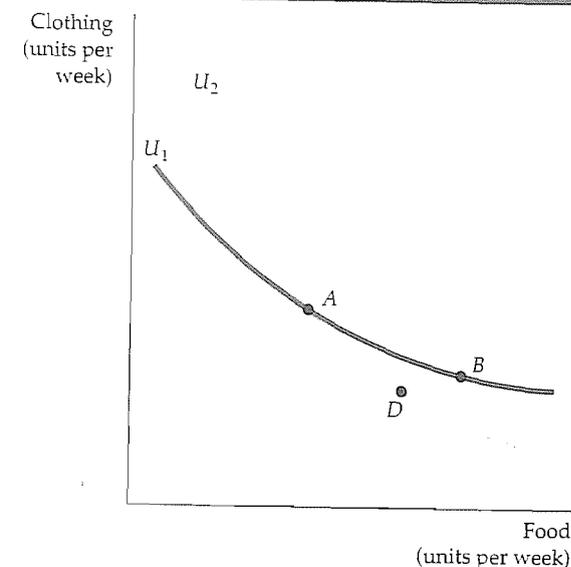


FIGURE 3.4 Indifference Curves Cannot Cross

If indifference curves U_1 and U_2 intersect, one of the assumptions of consumer theory is violated. According to this diagram, the consumer should be indifferent among market baskets A , B , and D . Yet B should be preferred to D because B has more of both goods.

contains more of both food and clothing. Thus, indifference curves that intersect contradicts our assumption that more is preferred to less.

Of course, there are an infinite number of nonintersecting indifference curves, one for every possible level of satisfaction. In fact, every possible market basket (each corresponding to a point on the graph) has an indifference curve passing through it.

The Shapes of Indifference Curves

Recall that indifference curves are all downward sloping. In our example of food and clothing, when the amount of food increases along an indifference curve, the amount of clothing decreases. The fact that indifference curves slope downward follows directly from our assumption that more of a good is better than less. If an indifference curve sloped upward, a consumer would be indifferent between two market baskets even though one of them had more of *both* food and clothing.

The shape of an indifference curve describes how a consumer is willing to substitute one good for another. As we saw in Chapter 1, people face trade-offs. The indifference curve in Figure 3.5 illustrates this principle. Starting at market basket A and moving to basket B , we see that the consumer is willing to give up 6 units of clothing to obtain 1 extra unit of food. However, in moving from B to D , he is willing to give up only 4 units of clothing to obtain an additional unit of food; in moving from D to E , he will give up only 2 units of clothing for 1 unit of food. The more clothing and the less food a person consumes, the more clothing he will give up in order to obtain more food. Similarly, the more food that a person possesses, the less clothing he will give up for more food.

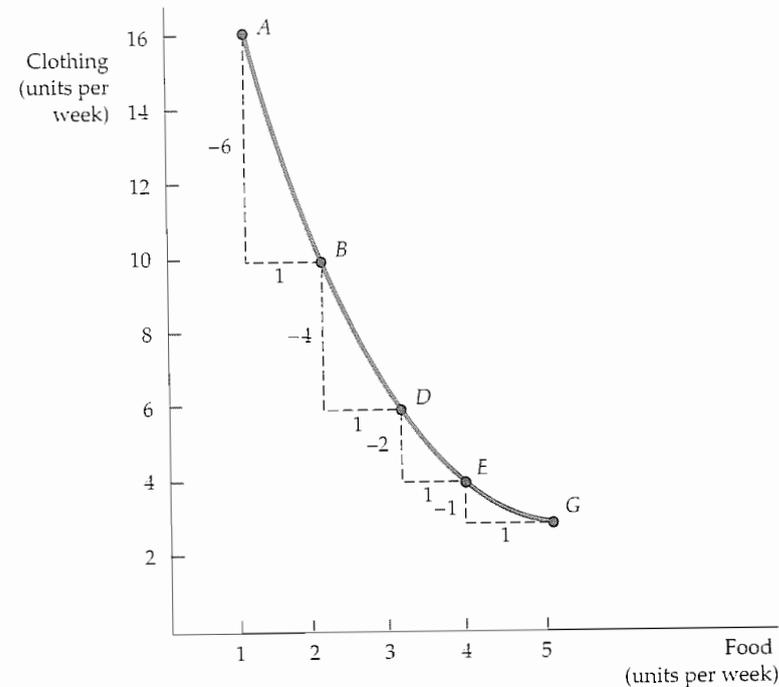


FIGURE 3.5 The Marginal Rate of Substitution

The slope of an indifference curve measures the consumer's marginal rate of substitution (MRS) between two goods. In this figure, the MRS between clothing (C) and food (F) falls from 6 (between A and B) to 4 (between B and D) to 2 (between D and E) to 1 (between E and G). When the MRS diminishes along an indifference curve, the curve is convex.

The Marginal Rate of Substitution

To quantify the amount of one good that a consumer will give up to obtain more of another, we use a measure called the **marginal rate of substitution (MRS)**. The MRS of food F for clothing C is the amount of clothing that a person is willing to give up to obtain one additional unit of food. Suppose, for example, the MRS is 3. This means that the consumer will give up 3 units of clothing to obtain 1 additional unit of food. If the MRS is 1/2, the consumer is willing to give up only 1/2 unit of clothing. Thus, the MRS measures the value that the individual places on 1 extra unit of one good in terms of another.

Look again at Figure 3.5. Note that clothing appears on the vertical axis and food on the horizontal axis. When we describe the MRS, we must be clear about which good we are giving up and which we are getting more of. To be consistent throughout the book, we will define the MRS in terms of the amount of the good on the vertical axis that the consumer is willing to give up to obtain 1 extra unit of the good on the horizontal axis. Thus in Figure 3.5, the MRS refers to the amount of clothing that the consumer is willing to give up to obtain an additional unit of food. If we denote the change in clothing by ΔC and the change in food by ΔF , the MRS can be written as $-\Delta C/\Delta F$. We add the negative sign to make the marginal rate of substitution a positive number (remember that ΔC is always negative; the consumer gives up clothing to obtain additional food).

marginal rate of substitution (MRS) Amount of a good that a consumer is willing to give up in order to obtain one additional unit of another good.

Thus the MRS at any point is equal in magnitude to the slope of the indifference curve. In Figure 3.5, for example, the MRS between points A and B is 6: The consumer is willing to give up 6 units of clothing to obtain 1 additional unit of food. Between points B and D, however, the MRS is 4: With these quantities of food and clothing, the consumer is willing to give up only 4 units of clothing to obtain 1 additional unit of food.

Convexity Also observe in Figure 3.5 that the MRS falls as we move down the indifference curve. This is not a coincidence. This decline in the MRS reflects an important characteristic of consumer preferences. To understand this, we will add an additional assumption regarding consumer preferences to the three that we discussed earlier in the chapter:

- 4. Diminishing marginal rate of substitution:** Indifference curves are *convex*, or bowed inward. The term *convex* means that the slope of the indifference curve *increases* (i.e., becomes less negative) as we move down along the curve. In other words, an indifference curve is convex if the MRS diminishes along the curve. The indifference curve in Figure 3.5 is convex. As we have seen, starting with market basket A in Figure 3.5 and moving to basket B, the MRS of food F for clothing C is $-\Delta C/\Delta F = -(-6)/1 = 6$. However, when we start at basket B and move from B to D, the MRS falls to 4. If we start at basket D and move to E, the MRS is 2. Starting at E and moving to G, we get an MRS of 1. As food consumption increases, the slope of the indifference curve falls in magnitude. Thus the MRS also falls.²

Is it reasonable to expect indifference curves to be convex? Yes. As more and more of one good is consumed, we can expect that a consumer will prefer to give up fewer and fewer units of a second good to get additional units of the first one. As we move down the indifference curve in Figure 3.5 and consumption of food increases, the additional satisfaction that a consumer gets from still more food will diminish. Thus, he will give up less and less clothing to obtain additional food.

Another way of describing this principle is to say that consumers generally prefer balanced market baskets to market baskets that contain all of one good and none of another. Note from Figure 3.5 that a relatively balanced market basket containing 3 units of food and 6 units of clothing (basket D) generates as much satisfaction as another market basket containing 1 unit of food and 16 units of clothing (basket A). It follows that a balanced market basket containing (for example) 6 units of food and 8 units of clothing will generate a higher level of satisfaction.

Perfect Substitutes and Perfect Complements

The shape of an indifference curve describes the willingness of a consumer to substitute one good for another. An indifference curve with a different shape implies a different willingness to substitute. To see this principle, look at the two polar cases illustrated in Figure 3.6.

Figure 3.6(a) shows Bob's preferences for apple juice and orange juice. These two goods are perfect substitutes for Bob because he is entirely indifferent between having a glass of one or the other. In this case, the MRS of apple juice for orange juice is 1: Bob is always willing to trade 1 glass of one for 1 glass of the

In §2.1, we explain that goods are *substitutes* when an increase in the price of one leads to an increase in the quantity demanded of the other.

² With nonconvex preferences, the MRS increases as the amount of the good measured on the horizontal axis increases along any indifference curve. This unlikely possibility might arise if one or both goods are addictive. For example, the willingness to substitute an addictive drug for other goods might increase as the use of the addictive drug increased.

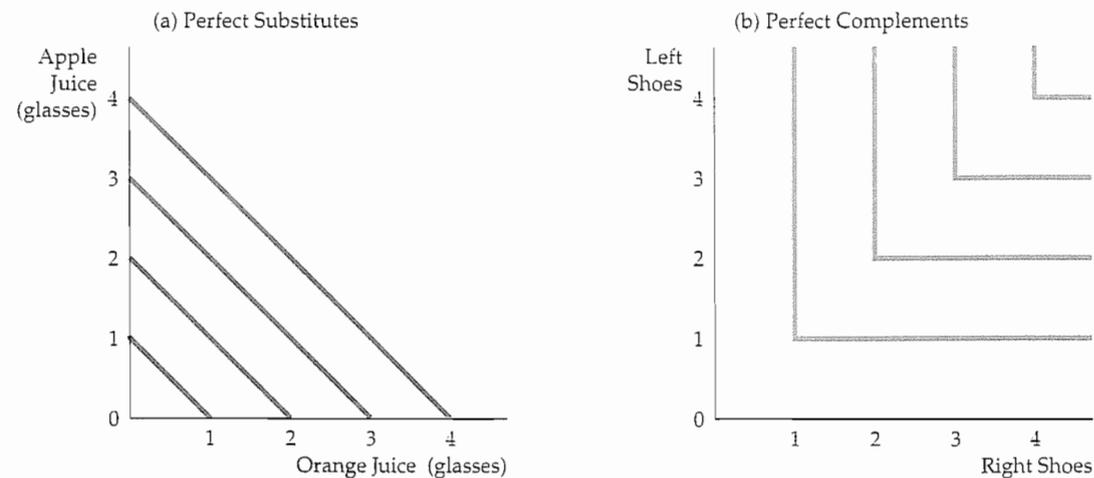


FIGURE 3.6 Perfect Substitutes and Perfect Complements

In (a), Bob views orange juice and apple juice as perfect substitutes: He is always indifferent between a glass of one and a glass of the other. In (b), Jane views left shoes and right shoes as perfect complements: An additional left shoe gives her no extra satisfaction unless she also obtains the matching right shoe.

perfect substitutes Two goods for which the marginal rate of substitution of one for the other is a constant.

In §2.1, we explain that goods are *complements* when an increase in the price of one leads to a decrease in the quantity demanded of the other.

perfect complements Two goods for which the MRS is infinite; the indifference curves are shaped as right angles.

bad Good for which less is preferred rather than more.

other. In general, we say that two goods are **perfect substitutes** when the marginal rate of substitution of one for the other is a constant. The indifference curves describing the trade-off between the consumption of the goods are straight lines. The slope of the indifference curves need not be -1 in the case of perfect substitutes. Suppose, for example, that Dan believes that one 16-megabyte memory chip is equivalent to two 8-megabyte chips because both combinations have the same memory capacity. In that case, the slope of Dan's indifference curve will be -2 (with the number of 8-megabyte chips on the vertical axis).

Figure 3.6(b) illustrates Jane's preferences for left shoes and right shoes. For Jane, the two goods are perfect complements because a left shoe will not increase her satisfaction unless she can obtain the matching right shoe. In this case, the MRS of left shoes for right shoes is zero whenever there are more right shoes than left shoes; Jane will not give up any left shoes to get additional right shoes. Correspondingly, the MRS is infinite whenever there are more left shoes than right because Jane will give up all but one of her excess left shoes in order to obtain an additional right shoe. Two goods are **perfect complements** when the indifference curves for both are shaped as right angles.

Bads So far, all of our examples have involved commodities that are "goods"—i.e., cases in which more of a commodity is preferred to less. However, some things are **bads**: *Less of them is preferred to more.* Air pollution is a bad; asbestos in housing insulation is another. How do we account for bads in the analysis of consumer preferences?

The answer is simple: We redefine the commodity under study so that the consumer tastes are represented as the preference for less of the bad. This reversal turns the bad into a good. Thus, for example, instead of a preference for air pollution, we will discuss the preference for clean air, which we can measure as

the degree of reduction in air pollution. Likewise, instead of referring to asbestos as a bad, we will refer to the corresponding good, the removal of asbestos.

With this simple adaptation, all four of the basic assumptions of consumer theory continue to hold, and we are ready to move on to an analysis of consumer budget constraints.

EXAMPLE 3.1 Designing New Automobiles (I)

If you were an automobile company executive, how would you decide when to introduce new models and how much money to invest in restyling? You would know that two of the most important attributes of a car are *styling* (e.g., design and interior features) and *performance* (e.g., gas mileage and handling). Both are desirable attributes: The better the styling and the performance, the greater will be the demand for a car. However, it costs money to restyle a car, and it also costs money to improve its performance. How much of each attribute should you include in your new model?

The answer depends in part on the costs of production, but it also depends on consumer preferences. Two characterizations of consumer preferences are shown in Figure 3.7. People with preferences shown in Figure 3.7(a) place greater value on performance than styling: They have a high MRS and are willing to give up quite a bit of styling to get better performance. Compare these preferences to those of a different segment of the population shown in Figure 3.7(b). These low-MRS people prefer styling to performance and will put up with poor gas mileage or handling to get a more stylish car.

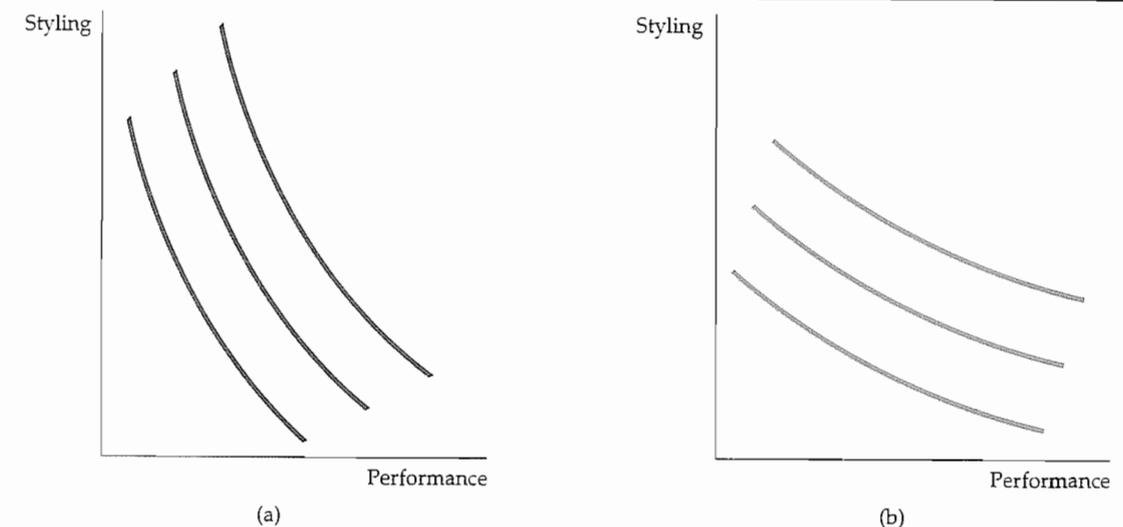


FIGURE 3.7 Preferences for Automobile Attributes

Preferences for automobile attributes can be described by indifference curves. Each curve shows the combinations of performance and styling that give the same satisfaction. Consumers in (a) are willing to give up a considerable amount of styling for additional performance. The opposite is true for consumers in (b).

Knowing which preference group is most prevalent can help executives make strategic production decisions. One way to obtain such information is by conducting surveys in which individuals are asked about their preferences for a number of automobiles with differing combinations of styling and performance. Another way is to analyze statistically past consumer purchases of cars that varied in styling and performance. By relating to their attributes the prices paid for different cars, we can determine the relative value attached to each attribute by various groups of consumers.³ Either approach can help determine whether the larger group more highly values performance (as in Figure 3.7a) or styling (as in Figure 3.7b). You can also determine the extent to which people in each group are willing to trade off one attribute for the other.

One study of automobile demand in the United States shows that over the past two decades, most consumers have preferred styling over performance.⁴ The study divided all cars sold in the United States into nine market classes, ranging from subcompact to luxury sport. Within each class, the degree of styling change was indexed from 1 (no visible exterior change) to 5 (a complete sheet metal change) to 9 (a completely new body, a change in size, and a conversion from rear-wheel to front-wheel drive). The study found that companies which emphasized style changes grew more rapidly than those that emphasized performance. In particular, cars undergoing major style changes enjoyed significantly higher sales growth than cars not undergoing such changes. (The major effect occurred immediately after the style change, but smaller effects were felt in subsequent years.)

The importance of styling helps explain the historic growth of Japanese imports in the United States: During the 1970s and 1980s, while U.S. domestic sales grew at 1.3 percent per year, import sales grew at 6.4 percent. On average, 15 percent of all domestic U.S. cars underwent a major style change each year, as compared to 23.4 percent of all imports. Although the market share of imports stabilized in the past decade, it is clear that styling changes (along with improvements in performance and reliability) spurred the growth of imported cars.

Utility You may have noticed a convenient feature of the theory of consumer behavior as we have described it so far: *It has not been necessary to associate a numerical level of satisfaction with each market basket consumed.* For example, with respect to the three indifference curves in Figure 3.3, we know that market basket A (or any other basket on indifference curve U_3) gives more satisfaction than any market basket on U_2 , such as B. Likewise, we know that the market baskets on U_2 are preferred to those on U_1 . The indifference curves simply allow us to describe consumer preferences graphically, building on the assumption that consumers can rank alternatives.

We will see that consumer theory relies only on the assumption that consumers can provide relative rankings of market baskets. Nonetheless, it is often useful to assign *numerical values* to individual baskets. Using this numerical

³ For an example, see Vladimir Bajic, "Automobiles and Implicit Markets: An Estimate of a Structural Demand Model for Automobile Characteristics," *Applied Economics* 25 (1993): 541–551.

⁴ See Edward L. Millner and George E. Hoffer, "A Reexamination of the Impact of Automotive Styling on Demand," *Applied Economics* 25 (1993): 101–110.

approach, we can describe consumer preferences by assigning scores to the levels of satisfaction associated with each indifference curve. In everyday language, the word *utility* has rather broad connotations, meaning, roughly, "benefit" or "well-being." Indeed, people obtain "utility" by getting things that give them pleasure and by avoiding things that give them pain. In the language of economics, the concept of **utility** refers to *the numerical score representing the satisfaction that a consumer gets from a market basket.* In other words, utility is a device used to simplify the ranking of market baskets. If buying three copies of this textbook makes you happier than buying one shirt, then we say that the books give you more utility than the shirt.

Utility Functions A **utility function** is a formula that assigns a level of utility to each market basket. Suppose, for example, that Phil's utility function for food (F) and clothing (C) is $u(F,C) = F + 2C$. In that case, a market basket consisting of 8 units of food and 3 units of clothing generates a utility of $8 + (2)(3) = 14$. Phil is therefore indifferent between this market basket and a market basket containing 6 units of food and 4 units of clothing ($6 + (2)(4) = 14$). On the other hand, either market basket is preferred to a third containing 4 units of food and 4 units of clothing. Why? Because this last market basket has a utility level of only $4 + (4)(2) = 12$.

We assign utility levels to market baskets so that if market basket A is preferred to basket B, the number will be higher for A than for B. For example, market basket A on the highest of three indifference curves U_3 might have a utility level of 3, while market basket B on the second-highest indifference curve U_2 might have a utility level of 2; on the lowest indifference curve U_1 , basket C, a utility level of 1. Thus the utility function provides the same information about preferences that an indifference map does: Both order consumer choices in terms of levels of satisfaction.

Let's examine one particular utility function in some detail. The *utility function* $u(F,C) = FC$ tells us that the level of satisfaction obtained from consuming F units of food and C units of clothing is the product of F and C . Figure 3.8 shows

utility Numerical score representing the satisfaction that a consumer gets from a given market basket.

utility function Formula that assigns a level of utility to individual market baskets.

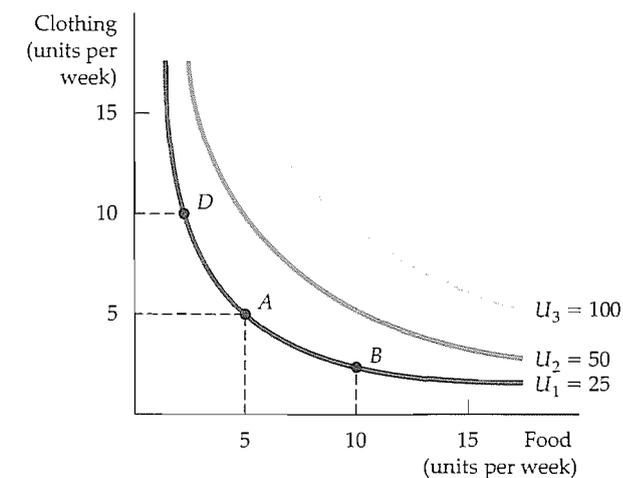


FIGURE 3.8 Utility Functions and Indifference Curves

A utility function can be represented by a set of indifference curves, each with a numerical indicator. This figure shows three indifference curves, with utility levels of 25, 50, and 100, respectively, associated with the utility function FC .

indifference curves associated with this function. The graph was drawn by initially choosing one particular market basket—say, $F = 5$ and $C = 5$ at point A . This market basket generates a utility level U_1 of 25. Then the indifference curve (also called an *isoutility curve*) was drawn by finding all market baskets for which $FC = 25$ (e.g., $F = 10, C = 2.5$ at point B ; $F = 2.5, C = 10$ at point D). The second indifference curve U_2 contains all market baskets for which $FC = 50$ and the third U_3 all market baskets for which $FC = 100$.

It is important to note that the numbers attached to the indifference curves are for convenience only. Suppose the utility function were changed to $u(F, C) = 4FC$. Consider any market basket that previously generated a utility level of 25—say, $F = 5$ and $C = 5$. Now the level of utility has increased, by a factor of 4, to 100. Thus the indifference curve labeled 25 looks the same, although it should now be labeled 100 rather than 25. In fact, the only difference between the indifference curves associated with the utility function $4FC$ and the utility function FC is that the curves are numbered 100, 200, and 400, rather than 25, 50, and 100. It is important to stress that the utility function is simply a way of *ranking* different market baskets; the magnitude of the utility difference between any two market baskets does not really tell us anything. The fact that U_3 has a level of utility of 100 and U_2 has a level of 50 does not mean that market baskets on U_3 generate twice as much satisfaction as those on U_2 . This is so because we have no means of objectively measuring a person's satisfaction or level of well-being from the consumption of a market basket. Thus whether we use indifference curves or a measure of utility, we know only that U_3 is better than U_2 and that U_2 is better than U_1 . We do not, however, know by *how much* one is preferred to the other.

Ordinal versus Cardinal Utility The three indifference curves in Figure 3.3 provide a ranking of market baskets that is ordered, or *ordinal*. For this reason, a utility function that generates a ranking of market baskets is called an **ordinal utility function**. The ranking associated with the ordinal utility function places market baskets in the order of most to least preferred. However, as explained above, it does not indicate by *how much* one is preferred to another. We know, for example, that any market basket on U_3 , such as A , is preferred to any on U_2 , such as B . However, the amount by which A is preferred to B (and B to D) is not revealed by the indifference map or by the ordinal utility function that generates it.

When working with ordinal utility functions, we must be careful to avoid a trap. Suppose that Juan's ordinal utility function attaches a utility level of 5 to a copy of this textbook; meanwhile Maria's utility function attaches a level of 10. Will Maria be happier than Juan if each of them gets a copy of this book? We don't know. Because these numerical values are arbitrary, interpersonal comparisons of utility are impossible.

When economists first studied utility and utility functions, they hoped that individual preferences could be quantified or measured in terms of basic units and could therefore provide a ranking that allowed for interpersonal comparisons. Using this approach, we could say that Maria gets twice as much satisfaction as Juan from a copy of this book. Or if we found that having a second copy increased Juan's utility level to 10, we could say that his happiness has doubled. If the numerical values assigned to market baskets did have meaning in this way, we would say that the numbers provided a *cardinal* ranking of alternatives. A utility function that describes by *how much* one market basket is preferred to another is called a **cardinal utility function**. Unlike ordinal utility functions, a

ordinal utility function
Utility function that generates a ranking of market baskets in order of most to least preferred.

cardinal utility function
Utility function describing by how much one market basket is preferred to another.

cardinal utility function attaches to market baskets numerical values that cannot arbitrarily be doubled or tripled without altering the differences between the values of various market baskets.

Unfortunately, we have no way of telling whether a person gets twice as much satisfaction from one market value as from another. Nor do we know whether one person gets twice as much satisfaction as another from consuming the same basket. (Could *you* tell whether you get twice as much satisfaction from consuming one thing versus another?) Fortunately, this constraint is unimportant. Because our objective is to understand consumer behavior, all that matters is knowing how consumers rank different baskets. Therefore, we will work only with ordinal utility functions. This approach is sufficient for understanding both how individual consumer decisions are made and what this knowledge implies about the characteristics of consumer demand.

3.2 Budget Constraints

So far we have focused only on the first piece of consumer theory—consumer preferences. We have seen how indifference curves (or, alternatively, utility functions) can be used to describe how consumers value various baskets of goods. Now we turn to the second part of consumer theory: the **budget constraints** that consumers face as a result of their limited incomes.

budget constraints Constraints that consumers face as a result of limited incomes.

The Budget Line

To see how a budget constraint limits a consumer's choices, let's consider a situation in which a woman has a fixed amount of income, I , that can be spent on food and clothing. Let F be the amount of food purchased and C the amount of clothing. We will denote the prices of the two goods P_F and P_C . In that case, P_FF (i.e., price of food times the quantity) is the amount of money spent on food and P_CC the amount of money spent on clothing.

The **budget line** indicates *all combinations of F and C for which the total amount of money spent is equal to income*. Because we are considering only two goods (and ignoring the possibility of saving), the woman will spend her entire income on food and clothing. As a result, the combinations of food and clothing that she can buy will all lie on this line:

$$P_FF + P_CC = I \quad (3.1)$$

budget line All combinations of goods for which the total amount of money spent is equal to income.

Suppose, for example, that our consumer has a weekly income of \$80, the price of food is \$1 per unit, and the price of clothing is \$2 per unit. Table 3.2 shows various combinations of food and clothing that she can purchase each week with her \$80. If her entire budget were allocated to clothing, the most that she could buy would be 40 units (at a price of \$2 per unit), as represented by market basket A . If she spent her entire budget on food, she could buy 80 units (at \$1 per unit), as given by market basket G . Market baskets B , D , and E show three additional ways in which \$80 could be spent on food and clothing.

MARKET BASKET	FOOD (F)	CLOTHING (C)	TOTAL SPENDING
A	0	40	\$80
B	20	30	\$80
D	40	20	\$80
E	60	10	\$80
G	80	0	\$80

Figure 3.9 shows the budget line associated with the market baskets given in Table 3.2. Because giving up a unit of clothing saves \$2 and buying a unit of food costs \$1, the amount of clothing given up for food along the budget line must be the same everywhere. As a result, the budget line is a straight line from point A to point G. In this particular case, the budget line is given by the equation $F + 2C = \$80$.

The intercept of the budget line is represented by basket A. As our consumer moves along the line from basket A to basket G, she spends less on clothing and more on food. It is easy to see that the extra clothing that must be given up to consume an additional unit of food is given by the ratio of the price of food to the price of clothing ($\$1/\$2 = 1/2$). Because clothing costs \$2 per unit and food only \$1 per unit, 1/2 unit of clothing must be given up to get 1 unit of food. In Figure 3.9 the slope of the line, $\Delta C/\Delta F = -1/2$, measures the relative cost of food and clothing.

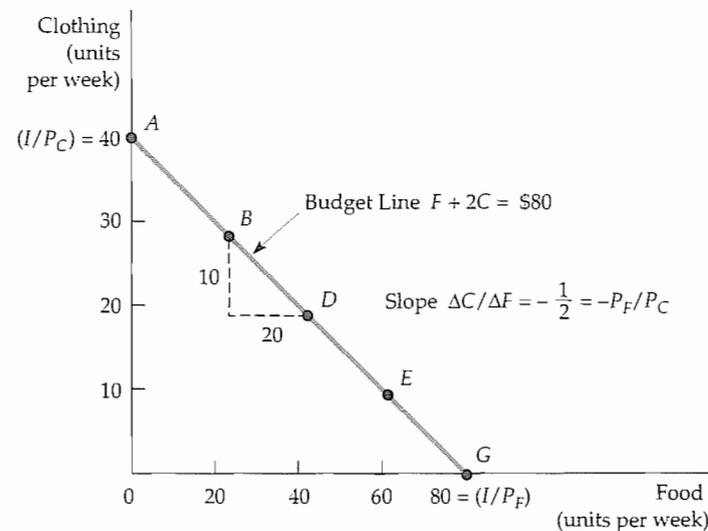


FIGURE 3.9 A Budget Line

A budget line describes the combinations of goods that can be purchased given the consumer's income and the prices of the goods. Line AG (which passes through points B, D, and E) shows the budget associated with an income of \$80, a price of food of $P_F = \$1$ per unit, and a price of clothing of $P_C = \$2$ per unit. The slope of the budget line (measured between points B and D) is $-P_F/P_C = -10/20 = -1/2$.

Using equation (3.1), we can see how much of C must be given up to consume more of F. We divide both sides of the equation by P_C and then solve for C:

$$C = (I/P_C) - (P_F/P_C)F \tag{3.2}$$

Equation (3.2) is the equation for a straight line; it has a vertical intercept of I/P_C and a slope of $-(P_F/P_C)$.

The slope of the budget line, $-(P_F/P_C)$, is the negative of the ratio of the prices of the two goods. The magnitude of the slope tells us the rate at which the two goods can be substituted for each other without changing the total amount of money spent. The vertical intercept (I/P_C) represents the maximum amount of C that can be purchased with income I. Finally, the horizontal intercept (I/P_F) tells us how many units of F can be purchased if all income were spent on F.

The Effects of Changes in Income and Prices

We have seen that the budget line depends both on income and on the prices of the goods P_F and P_C . But of course prices and income often change. Let's see how such changes affect the budget line.

Income Changes What happens to the budget line when income changes? From the equation for the straight line (3.2), we can see that a change in income alters the vertical intercept of the budget line but does not change the slope (because the price of neither good changed). Figure 3.10 shows that if income is doubled (from \$80 to \$160), the budget line shifts outward, from budget line L_1 to

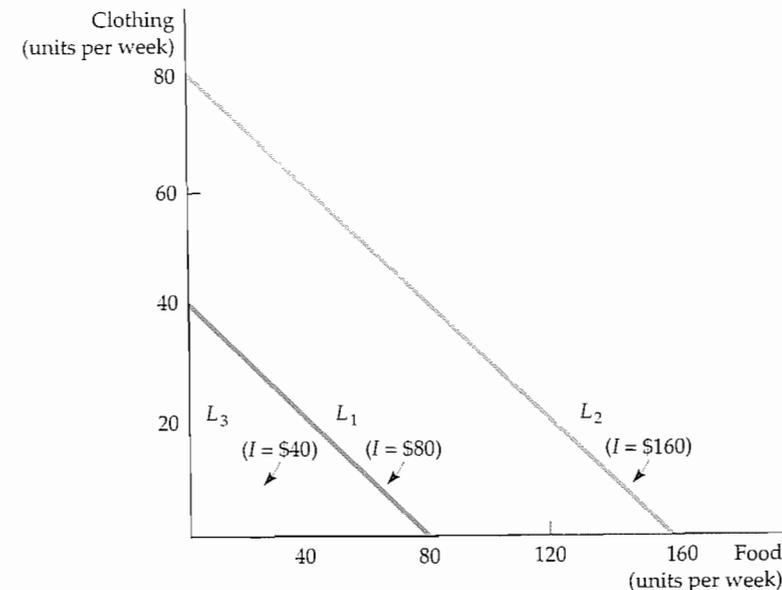


FIGURE 3.10 Effects of a Change in Income on the Budget Line

A change in income (with prices unchanged) causes the budget line to shift parallel to the original line (L_1). When the income of \$80 (on L_1) is increased to \$160, the budget line shifts outward to L_2 . If the income falls to \$40, the line shifts inward to L_3 .

budget line L_2 . Note, however, that L_2 remains parallel to L_1 . If she desires, our consumer can now double her purchases of both food and clothing. Likewise, if her income is cut in half (from \$80 to \$40), the budget line shifts inward, from L_1 to L_3 .

Price Changes What happens to the budget line if the price of one good changes but the price of the other does not? We can use the equation $C = (I/P_C) - (P_F/P_C)F$ to describe the effects of a change in the price of food on the budget line. Suppose the price of food falls by half, from \$1 to \$0.50. In that case, the vertical intercept of the budget line remains unchanged, although the slope changes from $-P_F/P_C = -1/\$2 = -1/2$ to $-\$0.50/\$2 = -1/4$. In Figure 3.11, we obtain the new budget line L_2 by rotating the original budget line L_1 outward, pivoting from the C-intercept. This rotation makes sense because a person who consumes only clothing and no food is unaffected by the price change. However, someone who consumes a large amount of food will experience an increase in his purchasing power. Because of the decline in the price of food, the maximum amount of food that can be purchased has doubled.

On the other hand, when the price of food doubles from \$1 to \$2, the budget line rotates inward to line L_3 because the person's purchasing power has diminished. Again, a person who consumed only clothing would be unaffected by the food price increase.

What happens if the prices of both food and clothing change, but in a way that leaves the *ratio* of the two prices unchanged? Because the slope of the budget line is equal to the ratio of the two prices, the slope will remain the same. The intercept of the budget line must shift so that the new line is parallel to the old one. For example, if the prices of both goods fall by half, then the slope of the budget line does not change. However, both intercepts double, and the budget line is shifted outward.

This exercise tells us something about the determinants of a consumer's *purchasing power*—her ability to generate utility through the purchase of goods and services. Purchasing power is determined not only by income, but also by

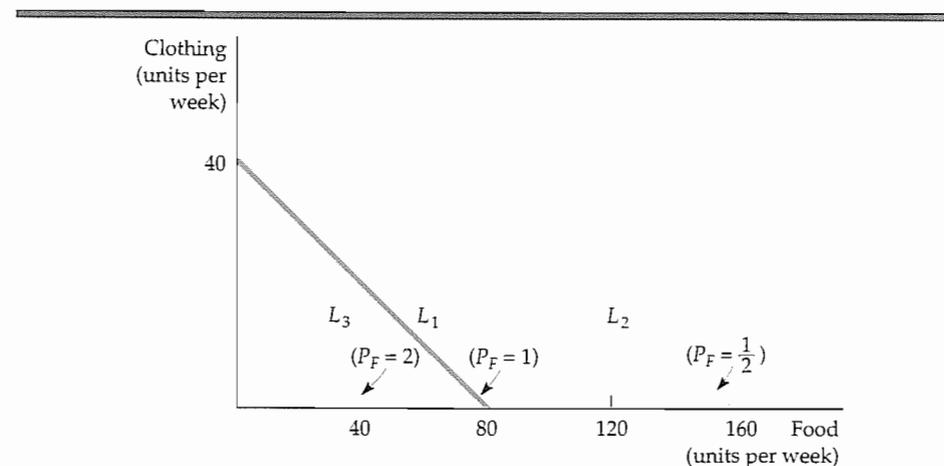


FIGURE 3.11 Effects of a Change in Price on the Budget Line

A change in the price of one good (with income unchanged) causes the budget line to rotate about one intercept. When the price of food falls from \$1.00 to \$0.50, the budget line rotates outward from L_1 to L_2 . However, when the price increases from \$1.00 to \$2.00, the line rotates inward from L_1 to L_3 .

prices. For example, our consumer's purchasing power can double either because her income doubles *or* because the prices of all the goods that she buys fall by half.

Finally, consider what happens if everything doubles—the prices of both food and clothing *and* the consumer's income. (This can happen in an inflationary economy.) Because both prices have doubled, the ratio of the prices has not changed; neither, therefore, has the slope of the budget line. Because the price of clothing has doubled along with income, the maximum amount of clothing that can be purchased (represented by the vertical intercept of the budget line) is unchanged. The same is true for food. Therefore, inflationary conditions in which all prices and income levels rise proportionately will not affect the consumer's budget line or purchasing power.

3.3 Consumer Choice

Given preferences and budget constraints, we can now determine how individual consumers choose how much of each good to buy. We assume that consumers make this choice in a rational way—that they choose goods to *maximize the satisfaction they can achieve, given the limited budget available to them*.

The maximizing market basket must satisfy two conditions:

1. *It must be located on the budget line.* To see why, note that any market basket to the left of and below the budget line leaves some income unallocated—income which, if spent, could increase the consumer's satisfaction. Of course, consumers can—and sometimes do—save some of their incomes for future consumption. In that case, the choice is not just between food and clothing, but between consuming food or clothing now and consuming food or clothing in the future. At this point, however, we will keep things simple by assuming that all income is spent now. Note also that any market basket to the right of and above the budget line cannot be purchased with available income. Thus, the only rational and feasible choice is a basket on the budget line.
2. *It must give the consumer the most preferred combination of goods and services.*

These two conditions reduce the problem of maximizing consumer satisfaction to one of picking an appropriate point on the budget line.

In our food and clothing example, as with any two goods, we can graphically illustrate the solution to the consumer's choice problem. Figure 3.12 shows how the problem is solved. Here, three indifference curves describe a consumer's preferences for food and clothing. Remember that of the three curves, the outermost curve U_3 , yields the greatest amount of satisfaction, curve U_2 the next greatest amount, and curve U_1 the least.

Note that point B on indifference curve U_1 is not the most preferred choice, because a reallocation of income in which more is spent on food and less on clothing can increase the consumer's satisfaction. In particular, by moving to point A , the consumer spends the same amount of money and achieves the increased level of satisfaction associated with indifference curve U_2 . In addition, note that baskets located to the right and above indifference curve U_2 , like the basket associated with D on indifference curve U_3 , achieve a higher level of satisfaction but cannot be purchased with the available income. Therefore, A maximizes the consumer's satisfaction.

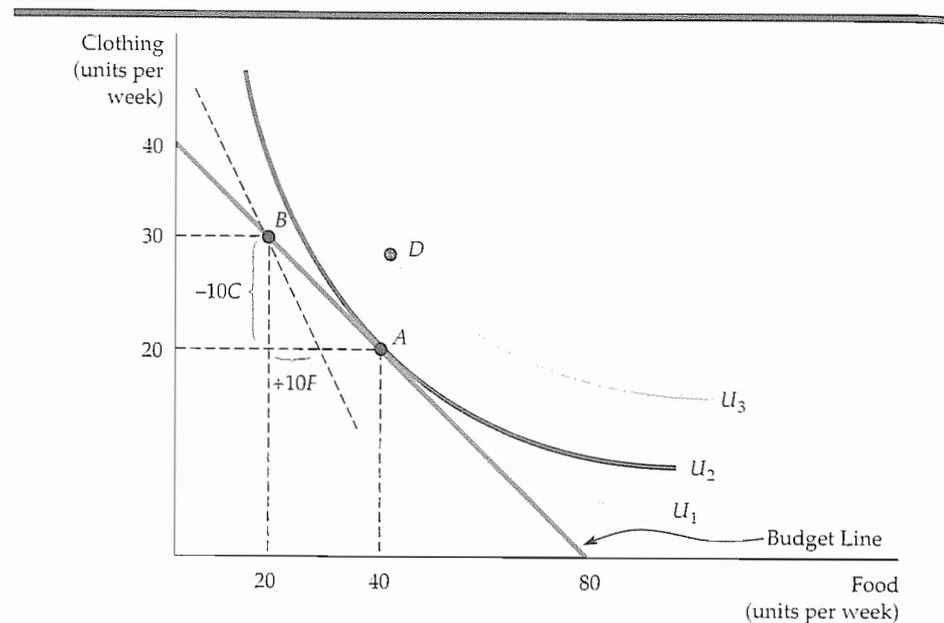


FIGURE 3.12 Maximizing Consumer Satisfaction

Consumers maximize satisfaction by choosing market basket A. At this point, the budget line and indifference curve U_2 are tangent, and no higher level of satisfaction (e.g., with market basket D) can be attained. At A, the point of maximization, the MRS between the two goods equals the price ratio. At B, however, because the MRS [$-(-10/10) = 1$] is greater than the price ratio ($1/2$), satisfaction is not maximized.

We see from this analysis that the basket which maximizes satisfaction must lie on the highest indifference curve that touches the budget line. Point A is the point of tangency between indifference curve U_2 and the budget line. At A, the slope of the budget line is exactly equal to the slope of the indifference curve. Because the MRS ($-\Delta C/\Delta F$) is the negative of the slope of the indifference curve, we can say that satisfaction is maximized (given the budget constraint) at the point where

$$\text{MRS} = P_F/P_C \quad (3.3)$$

This is an important result: Satisfaction is maximized when the *marginal rate of substitution* (of F for C) is equal to the ratio of the prices (of F to C). Thus the consumer can obtain maximum satisfaction by adjusting his consumption of goods F and C so that the MRS equals the price ratio.

The condition given in equation (3.3) illustrates the kinds of optimization conditions that arise in economics. In this instance, satisfaction is maximized when the **marginal benefit**—the benefit associated with the consumption of one additional unit of food—is equal to the **marginal cost**—the cost of the additional unit of food. The marginal benefit is measured by the MRS. At point A, it equals $1/2$ (the magnitude of the slope of the indifference curve), which implies that the consumer is willing to give up $1/2$ unit of clothing to obtain 1 unit of food. At the same point, the marginal cost is measured by the magnitude of the slope of the budget line; it too equals $1/2$ because the cost of getting one unit of food is giving up $1/2$ unit of clothing ($P_F = 1$ and $P_C = 2$ on the budget line).

marginal benefit Benefit from the consumption of one additional unit of a good.

marginal cost Cost of one additional unit of a good.

If the MRS is less or greater than the price ratio, the consumer's satisfaction has not been maximized. For example, compare point B in Figure 3.12 to point A. At point B, the consumer is purchasing 20 units of food and 30 units of clothing. The price ratio (or marginal cost) is equal to $1/2$ because food costs \$1 and clothing \$2. However, the MRS (or marginal benefit) is greater than $1/2$; it is approximately 1. As a result, the consumer is able to substitute 1 unit of food for 1 unit of clothing without loss of satisfaction. Because food is cheaper than clothing, it is in her interest to buy more food and less clothing. If our consumer purchases 1 less unit of clothing, for example, the \$2 saved can be allocated to two units of food even though only one unit is needed to maintain her level of satisfaction.⁵

The reallocation of the budget continues in this manner (moving along the budget line), until we reach point A, where the price ratio of $1/2$ just equals the MRS of $1/2$. This point implies that the consumer is willing to trade one unit of clothing for two units of food. Only when the condition $\text{MRS} = 1/2 = P_F/P_C$ holds is she maximizing her satisfaction.

EXAMPLE 3.2 Designing New Automobiles (II)

Our analysis of consumer choice allows us to see how the differing preferences of consumer groups for automobiles can affect their purchasing decisions. Following up on Example 3.1, we consider two groups of consumers. The members of each group wish to spend \$10,000 each on the styling and performance of a new car. (Additional money could be allocated to other attributes of automobiles not discussed here; thus the total expenditure on each car could be more than \$10,000.) Each group has different preferences for styling and performance.

Figure 3.13 shows the car-buying budget constraint faced by individuals in each group. The first group, with preferences similar to those in Figure 3.7(a), prefers performance to styling. By finding the point of tangency between a typical individual's indifference curve and the budget constraint, we see that consumers in this group would prefer to buy a car whose performance was worth \$7,000 and whose styling was worth \$3,000. Individuals in the second group, however, would prefer cars with \$2,500 worth of performance and \$7,500 worth of styling. (Recall from Example 3.1 that statistical studies have shown that the majority of consumers belong to the second group.)

With knowledge of group preferences, an automobile company can design a production and marketing plan. One potentially profitable option is to appeal to both groups by manufacturing a model emphasizing styling to a slightly lesser degree than preferred by individuals in Figure 3.13(b) but to a much greater degree than preferred by those in Figure 3.13(a). A second option is to produce a relatively large number of cars that emphasize styling and a smaller number emphasizing performance. Knowledge about the preferences of each group, along with information about the number of consumers in each,

⁵ The result that the MRS equals the price ratio is deceptively powerful. Imagine two consumers who have just purchased various quantities of food and clothing. Without looking at their purchases, you can tell both persons (if they are maximizing) the value of their MRS (by looking at the prices of the two goods). What you cannot tell, however, is the quantity of each good purchased, because that decision is determined by their individual preferences. If the two consumers have different tastes, they will consume different quantities of food and clothing, even though each MRS is the same.

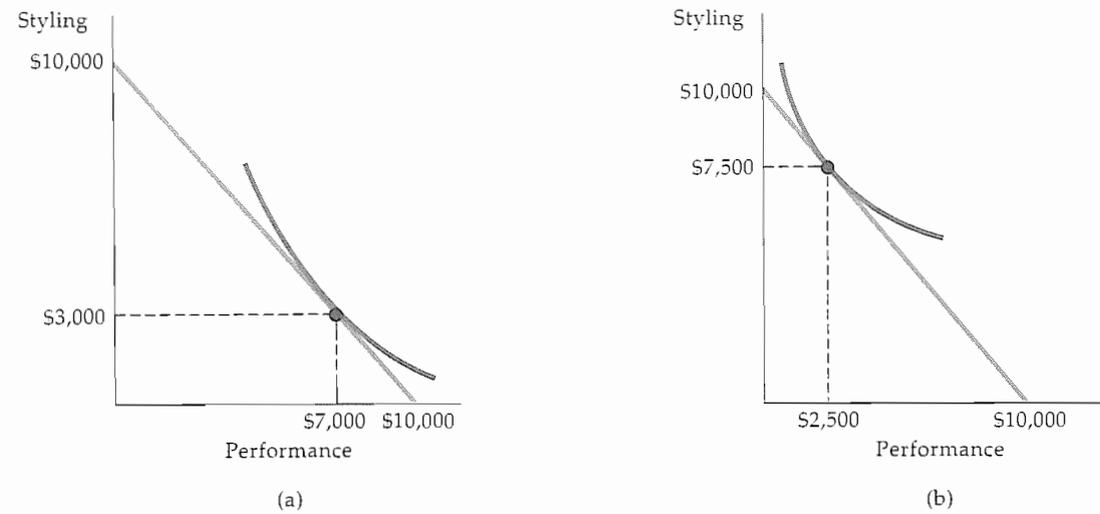


FIGURE 3.13 Consumer Choice of Automobile Attributes

The consumers in (a) are willing to trade off a considerable amount of styling for some additional performance. Given a budget constraint, they will choose a car that emphasizes performance. The opposite is true for consumers in (b).

would be sufficient to allow the firm to make a sensible strategic business decision.

In fact, an exercise similar to this was carried out by General Motors in a survey of a large number of automobile buyers.⁶ Some of the results were expected. For example, households with children tended to prefer functionality over style and so tended to buy minivans rather than sedans and sporty cars. Rural households, on the other hand, tended to purchase pickups and all-wheel drives. More interesting was the strong correlation between age and preferences for attributes. Older consumers tended to prefer larger and heavier cars with more safety features and accessories (e.g., power windows and steering). Younger consumers preferred greater horsepower and more stylish cars (including sport utility vehicles).

EXAMPLE 3.3 Decision Making and Public Policy

Grant programs from the federal government to state and local governments serve many purposes. One program might seek to increase school spending, another to redistribute income from relatively wealthy states and localities to those that are relatively poor. A third might try to ensure that individual governments provide minimum service levels.

⁶ The survey design and the results are described in Steven Berry, James Levinsohn, and Ariel Pakes, "Differentiated Products Demand Systems from a Combination of Micro and Macro Data: The New Car Market," National Bureau of Economic Research Working Paper 6481, March 1998.

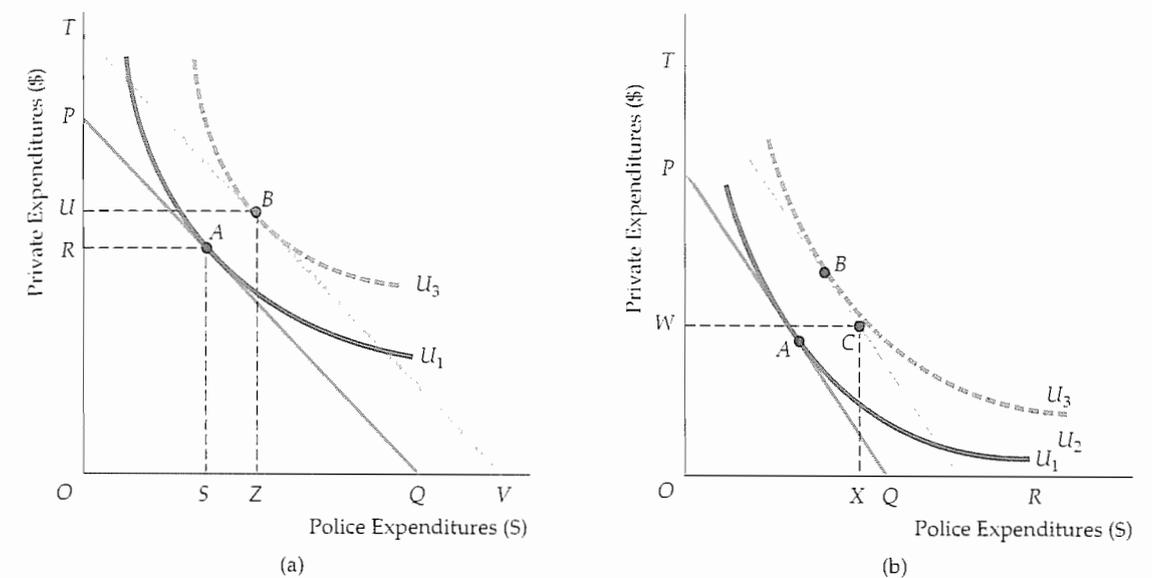


FIGURE 3.14 (a) A Nonmatching Grant (b) A Matching Grant

(a) A nonmatching grant from the federal government to a local government acts just like an income increase in traditional consumer analysis. The local government official moves from A to B, allocating a portion of the grant to police expenditures and a portion to lower taxes and, therefore, to an increase in private expenditures. (b) A matching grant acts just like a price decrease in traditional consumer analysis. The local government official moves from A to C, allocating some of the grant to police expenditures and some to private expenditures. Relatively more money, however, is spent on police expenditures than would be the case with a nonmatching grant of the same total amount.

Which kinds of grant programs are best suited to achieve these different objectives? The answer depends on the incentive effects that each program generates. By changing the constraints faced by local public officials, a grant program can alter an official's decision about how much a local government should spend. We can use consumer theory to see how two types of grant programs evoke different responses from public officials.

Suppose that a public official is in charge of the police budget, which is paid for by local taxes. Her preferences reflect what she believes should be allocated for police spending and what she feels citizens would prefer to have available for private consumption. Before the introduction of the grant program, the city's budget line is PQ in Figure 3.14(a). This budget line represents the total amount of resources available for public police spending (shown on the horizontal axis) and private spending (on the vertical).⁷ The preference-maximizing market basket A on indifference curve U_1 shows that OR is spent on private expenditures and OS on police expenditures. Because public expenditures are paid for by local taxes, these private expenditures represent spending after local taxes have been paid.

The first type of grant program, a *nonmatching grant*, is simply a check from the federal government that the local government can spend without restriction. An unconditional grant of this sort expands the community budget line

⁷ This sum would approximately equal the per capita income of the jurisdiction (say \$10,000) times the number of taxpayers (say, 50,000).

outward from PQ to TV in Figure 3.14(a), where $PT = QV$ is the dollar amount of the grant. The local government's response to this influx of dollars is to move to a higher indifference curve by selecting market basket B , with more of both goods (OU of private expenditures and OZ of police expenditures). But more private expenditures means that some of the money for police that came previously from taxes now comes from government grants.

The second type of program is the *matching grant*—funds offered as a form of subsidy to local spending. For example, the federal government might offer \$1 for every \$2 that the local government raises to pay for police. As a result, a matching grant lowers the cost of the publicly provided good. In terms of Figure 3.14(b), the matching grant rotates the budget line outward from PQ to PR . If no local money is spent on police, the budget line is unchanged. However, if the local official decides to spend money on the public sector, the budget increases.

In response to the matching grant, the official chooses market basket C rather than A . As with a nonmatching grant, there is an increase in police expenditures and a tax cut that leads to an increase in private consumption. At C , OX dollars are allocated to police and OW to private expenditures. However, the spending effects of the two types of grant are different. The diagram shows that the matching grant leads to greater police spending than does the nonmatching grant, even when the two programs involve identical government expenditures.⁸

Corner Solutions

Sometimes consumers buy in extremes, at least within categories of goods. Some people, for example, spend no money on travel and entertainment. Indifference curve analysis can be used to show conditions under which consumers choose not to consume a particular good.

In Figure 3.15, a man faced with budget line for snacks AB chooses to purchase only ice cream (IC) and no frozen yogurt (Y). This decision reflects what is called a **corner solution**: When one of the goods is not consumed, the consumption bundle appears at the corner of the graph. At B , which is the point of maximum satisfaction, the MRS of ice cream for frozen yogurt is greater than the slope of the budget line. This inequality suggests that if the consumer had more frozen yogurt to give up, he would gladly trade it for additional ice cream. At this point, however, our consumer is already consuming all ice cream and no frozen yogurt, and it is impossible to consume *negative* amounts of frozen yogurt.

When a corner solution arises, the consumer's MRS does not necessarily equal the price ratio. Unlike the condition expressed in equation (3.3), the necessary condition for satisfaction to be maximized when choosing between ice cream and frozen yogurt in a corner solution is given by the following inequality:⁹

$$\text{MRS} \geq P_{IC}/P_Y \quad (3.4)$$

⁸ Note also that point B , which is attained with a nonmatching grant, is on a higher indifference curve than point C , which is attained with a matching grant. The nonmatching grant leads to greater satisfaction for the same level of expenditure. In other words, there is a trade-off between encouraging a particular change in expenditure and achieving the highest level of satisfaction for a given expenditure.

⁹ Strict equality could hold if the slope of the budget constraint happened to equal the slope of the indifference curve—a condition that is unlikely.

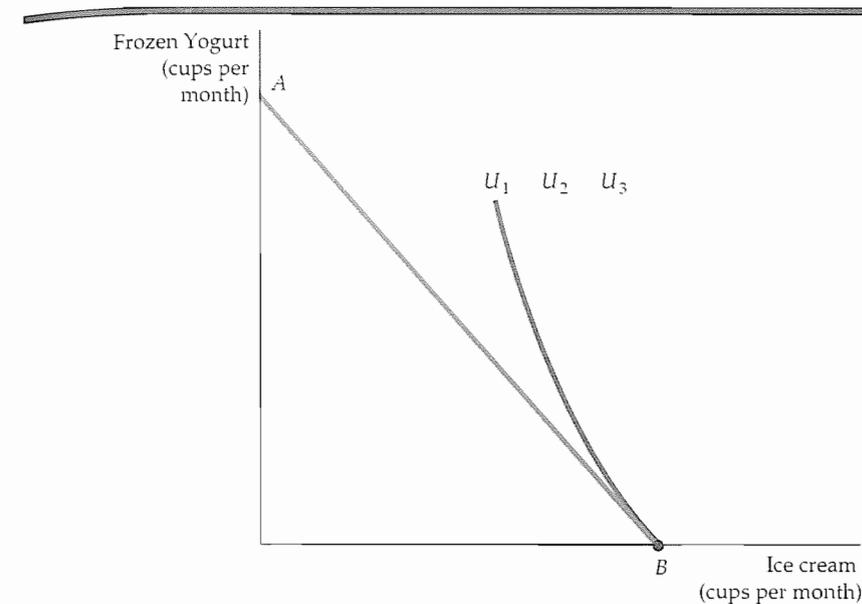


FIGURE 3.15 A Corner Solution

When the consumer's marginal rate of substitution is not equal to the price ratio for all levels of consumption, a corner solution arises. The consumer maximizes satisfaction by consuming only one of the two goods. Given budget line AB , the highest level of satisfaction is achieved at B on indifference curve U_1 , where the MRS (of ice cream for frozen yogurt) is greater than the ratio of the price of ice cream to the price of frozen yogurt.

This inequality would, of course, be reversed if the corner solution were at point A rather than B . In either case, we can see that the marginal benefit–marginal cost equality that we described in the previous section holds only when positive quantities of all goods are consumed.

An important lesson here is that predictions about how much of a product consumers will purchase when faced with changing economic conditions depend on the nature of consumer preferences for that product and related products and on the slope of the consumer's budget line. If the MRS of ice cream for frozen yogurt is substantially greater than the price ratio, as in Figure 3.15, then a small decrease in the price of frozen yogurt will not alter the consumer's choice; he will still choose to consume only ice cream. But if the price of frozen yogurt falls far enough, the consumer could quickly choose to consume a lot of frozen yogurt.

EXAMPLE 3.4 A College Trust Fund

Jane Doe's parents have provided a trust fund for her college education. Jane, who is 18, can receive the entire trust fund on the condition that she spend it only on education. The fund is a welcome gift to Jane but perhaps not as welcome as an unrestricted trust. To see why Jane feels this way, consider Figure 3.16, in which dollars per year spent on education are shown on the horizontal axis and dollars spent on other forms of consumption on the vertical.

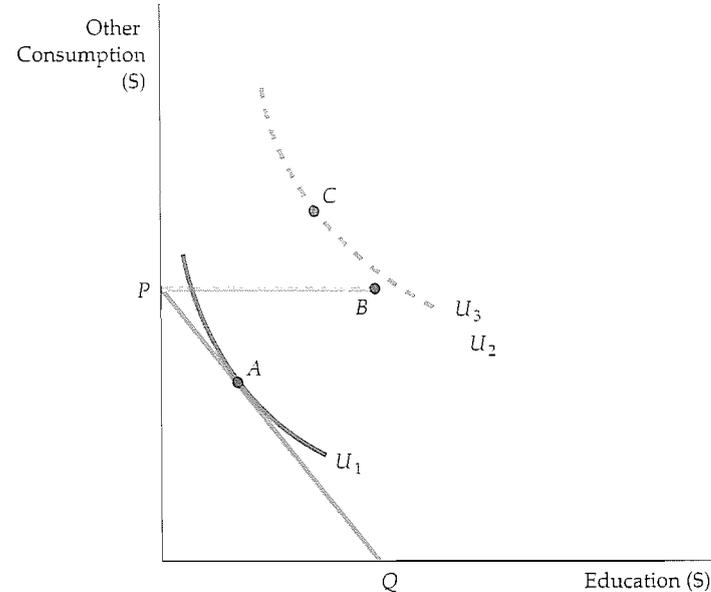


FIGURE 3.16 A College Trust Fund

When given a college trust fund that must be spent on education, the student moves from A to B , a corner solution. If, however, the trust fund could be spent on other consumption as well as education, the student would be better off at C .

The budget line that Jane faces before being awarded the trust is given by line PQ . The trust fund expands the budget line outward as long as the full amount of the fund, shown by distance PB , is spent on education. By accepting the trust fund and going to college, Jane increases her satisfaction, moving from A on indifference curve U_1 to B on indifference curve U_2 .

Note that B represents a corner solution because Jane's marginal rate of substitution of other consumption for education is lower than the relative price of other consumption. Jane would prefer to spend a portion of the trust fund on other goods in addition to education. Without restriction on the trust fund, she would move to C on indifference curve U_3 , decreasing her spending on education (perhaps going to a junior college rather than a four-year college) but increasing her spending on items that she enjoys more than education.

Recipients usually prefer unrestricted to restricted trusts. Restricted trusts are popular, however, because they allow parents to control children's expenditures in ways that they believe are in the children's long-run best interests.

3.4 Revealed Preference

In Section 3.1, we saw how an individual's preferences could be represented by a series of indifference curves. Then in Section 3.3, we saw how preferences, given budget constraints, determine choices. Can this process be reversed? If we know the choices that a consumer has made, can we determine his or her preferences?

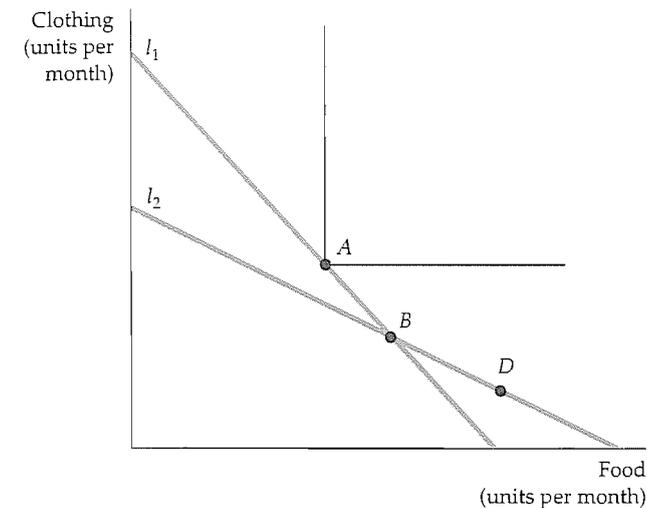


FIGURE 3.17 Revealed Preference: Two Budget Lines

If an individual facing budget line l_1 has chosen market basket A rather than market basket B , A is revealed to be preferred to B . Likewise, the individual facing budget line l_2 chooses market basket B , which is then revealed to be preferred to market basket D . Whereas A is preferred to all market baskets in the green-shaded area, all baskets in the pink-shaded area are preferred to A .

We can if we have information about a sufficient number of choices that have been made when prices and income levels varied. The basic idea is simple. *If a consumer chooses one market basket over another, and if the chosen market basket is more expensive than the alternative, then the consumer must prefer the chosen market basket.*

Suppose that an individual, facing the budget constraint given by line l_1 in Figure 3.17, chooses market basket A . Let's compare A to baskets B and D . Because the individual could have purchased basket B (and all baskets below line l_1) and did not, we say that A is preferred to B .

It might seem at first glance that we cannot make a direct comparison between baskets A and D because D is not on l_1 . But suppose the relative prices of food and clothing change, so that the new budget line is l_2 and the individual then chooses market basket B . Because D lies on budget line l_2 and was not chosen, B is preferred to D (and to all baskets below line l_2). Because A is preferred to B and B is preferred to D , we conclude that A is preferred to D . Furthermore, note in Figure 3.17 that basket A is preferred to all of the baskets that appear in the green-shaded areas. However, because food and clothing are "goods" rather than "bads," all baskets that lie in the pink-shaded area in the rectangle above and to the right of A are preferred to A . Thus, the indifference curve passing through A must lie in the unshaded area.

Given more information about choices when prices and income levels vary, we can get a better fix on the shape of the indifference curve. Consider Figure 3.18. Suppose that facing line l_3 (which was chosen to pass through A), the individual chooses market basket E . Because E was chosen even though A was equally expensive (it lies on the same budget line), E is preferred to A , as are all points in the rectangle above and to the right of E . Now suppose that facing line l_4 (which passes through A), the individual chooses market basket G . Because G was chosen and A was not, G is preferred to A , as are all market baskets above and to the right of G .

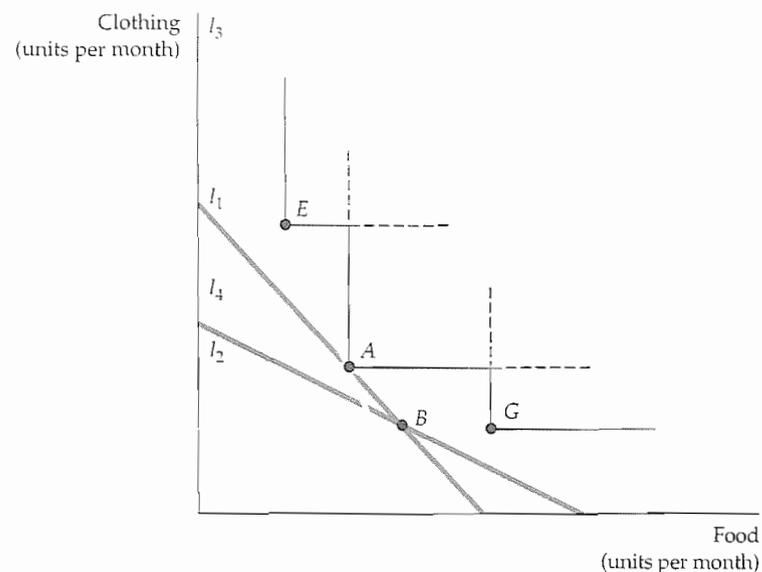


FIGURE 3.18 Revealed Preference: Four Budget Lines

Facing budget line l_3 the individual chooses E , which is revealed to be preferred to A (because A could have been chosen). Likewise, facing line l_4 , G is chosen, which is also revealed to be preferred to A . Whereas A is preferred to all market baskets in the green-shaded area, all market baskets in the pink-shaded area are preferred to A .

We can go further by making use of the assumption that preferences are convex. In that case, because E is preferred to A , all market baskets above and to the right of line AE in Figure 3.18 must be preferred to A . Otherwise, the indifference curve passing through A would have to pass through a point above and to the right of AE and then fall below the line at E —in which case the indifference curve would not be convex. By a similar argument, all points on AG or above are also preferred to A . Therefore, the indifference curve must lie within the unshaded area.

The revealed preference approach is valuable as a means of checking whether individual choices are consistent with the assumptions of consumer theory. As Example 3.5 shows, revealed preference analysis can help us understand the implications of choices that consumers must make in particular circumstances.

EXAMPLE 3.5 Revealed Preference for Recreation

A health club has been offering the use of its facilities to anyone who is willing to pay an hourly fee. Now the club decides to alter its pricing policy by charging both an annual membership fee and a lower hourly fee. Does this new financial arrangement make individuals better off or worse off than they were under the old arrangement? The answer depends on people's preferences.

Suppose that Roberta has \$100 of income available each week for recreational activities, including exercise, movies, restaurant meals, and so on. When the health club charged a fee of \$4 per hour, Roberta used the facility 10 hours per week. Under the new arrangement, she is required to pay \$30 per week but can use the club for only \$1 per hour.

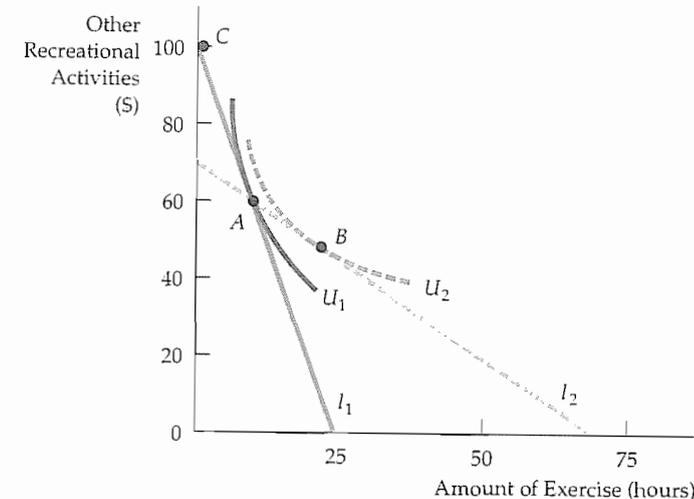


FIGURE 3.19 Revealed Preference for Recreation

When facing budget line l_1 , an individual chooses to use a health club for 10 hours per week at point A . When the fees are altered, she faces budget line l_2 . She is then made better off because market basket A can still be purchased, as can market basket B , which lies on a higher indifference curve.

Is this change beneficial for Roberta? Revealed preference analysis provides the answer. In Figure 3.19, line l_1 represents the budget constraint that Roberta faced under the original pricing arrangement. In this case she maximized her satisfaction by choosing market basket A , with 10 hours of exercise and \$60 of other recreational activities. Under the new arrangement, which shifts the budget line to l_2 , she could still choose market basket A . But because U_1 is clearly not tangent to l_2 , Roberta will be better off choosing another basket, such as B , with 25 hours of exercise and \$45 of other recreational activities. Because she would choose B when she could still choose A , she prefers B to A . The new pricing arrangement therefore makes Roberta better off. (Note that B is also preferred to C , which represents the option of not using the health club at all.)

We could also ask whether this new pricing system—called a *two-part tariff*—will increase the club's profits. If all members are like Roberta and more use generates more profit, then the answer is yes. In general, however, the answer depends on two factors: the preferences of all members and the costs of operating the facility. We discuss the two-part tariff in detail in Chapter 11, where we study ways in which firms with market power set prices.

3.5 Marginal Utility and Consumer Choice

In Section 3.3, we showed graphically how a consumer can maximize his or her satisfaction given a budget constraint. We do this by finding the highest indifference curve that can be reached, given that budget constraint. Because the highest

indifference curve also has the highest attainable level of utility, it is natural to recast the consumer's problem as one of maximizing utility subject to a budget constraint.

The concept of utility can also be used to recast our analysis in a way that provides additional insight. To begin, let's distinguish between the total utility obtained by consumption and the satisfaction obtained from the last item consumed. **Marginal utility (MU)** measures *the additional satisfaction obtained from consuming one additional unit of a good*. For example, the marginal utility associated with a consumption increase from 0 to 1 unit of food might be 9; from 1 to 2, it might be 7; from 2 to 3, it might be 5.

marginal utility (MU) Additional satisfaction obtained from consuming one additional unit of a good.

diminishing marginal utility Principle that as more of a good is consumed, the consumption of additional amounts will yield smaller additions to utility.

These numbers imply that the consumer has **diminishing marginal utility**: As more and more of a good is consumed, consuming additional amounts will yield smaller and smaller additions to utility. Imagine, for example, the consumption of television: Marginal utility might fall after the second or third hour and could become very small after the fourth or fifth.

We can relate the concept of marginal utility to the consumer's utility-maximization problem in the following way. Consider a small movement down an indifference curve in Figure 3.8 (p. 73). The additional consumption of food, ΔF , will generate marginal utility MU_F . This shift results in a total increase in utility of $MU_F \Delta F$. At the same time, the reduced consumption of clothing, ΔC , will lower utility per unit by MU_C , resulting in a total loss of $MU_C \Delta C$.

Because all points on an indifference curve generate the same level of utility, the total gain in utility associated with the increase in F must balance the loss due to the lower consumption of C . Formally,

$$0 = MU_F(\Delta F) + MU_C(\Delta C)$$

Now we can rearrange this equation so that

$$-(\Delta C/\Delta F) = MU_F/MU_C$$

But because $-(\Delta C/\Delta F)$ is the MRS of F for C , it follows that

$$\text{MRS} = MU_F/MU_C \quad (3.5)$$

Equation (3.5) tells us that the MRS is the ratio of the marginal utility of F to the marginal utility of C . As the consumer gives up more and more of C to obtain more of F , the marginal utility of F falls and that of C increases.

We saw earlier in this chapter that when consumers maximize their satisfaction, the MRS of F for C is equal to the ratio of the prices of the two goods:

$$\text{MRS} = P_F/P_C \quad (3.6)$$

Because the MRS is also equal to the ratio of the marginal utilities of consuming F and C (from equation 3.5), it follows that

$$MU_F/MU_C = P_F/P_C$$

or

$$MU_F/P_F = MU_C/P_C \quad (3.7)$$

Equation (3.7) is an important result. It tells us that utility maximization is achieved when the budget is allocated so that *the marginal utility per dollar of expenditure is the same for each good*. To see why this principle must hold, suppose that a person gets more utility from spending an additional dollar on food than on clothing. In this case, her utility will be increased by spending more on food. As long as the marginal utility of spending an extra dollar on food exceeds the marginal utility of spending an extra dollar on clothing, she can increase her utility by shifting her budget toward food and away from clothing. Eventually, the marginal utility of food will decrease (because there is diminishing marginal utility in its consumption) and the marginal utility of clothing will increase (for the same reason). Only when the consumer has satisfied the **equal marginal principle**—i.e., *has equalized the marginal utility per dollar of expenditure across all goods*—will she have maximized utility. The equal marginal principle is an important concept in microeconomics. It will reappear in different forms throughout our analysis of consumer and producer behavior.

equal marginal principle Principle that utility is maximized when the consumer has equalized the marginal utility per dollar of expenditure across all goods.

EXAMPLE 3.6 Gasoline Rationing

In times of war and other crises, governments often impose price controls on critical products. In 1974 and 1979, for example, the U.S. government imposed price controls on gasoline. As a result, motorists wanted to buy more gasoline than was available at controlled prices, and gasoline had to be rationed. Nonprice rationing is an alternative way of dealing with shortages that some people consider fairer than relying on uncontested market forces. Under one form of rationing, everyone has an equal chance to purchase a rationed good. Under a market system, those with higher incomes can outbid those with lower incomes to obtain goods that are in scarce supply.

In the United States, gasoline was allocated by long lines at the gas pumps: While those who were willing to give up their time waiting got the gas they wanted, others did not. By guaranteeing every eligible person a minimum amount of gasoline, rationing can provide some people with access to a product that they could not otherwise afford. But rationing hurts others by limiting the amount of gasoline that they can buy.¹⁰

We can see this principle clearly in Figure 3.20, which applies to a woman with an annual income of \$20,000. The horizontal axis shows her annual consumption of gasoline, the vertical axis her remaining income after purchasing gasoline. Suppose the controlled gasoline price is \$1 per gallon. Because her income is \$20,000, she is limited to the points on budget line AB , which has a slope of -1 . At \$1 per gallon, she might wish to buy 5,000 gallons of gasoline per year and spend \$15,000 on other goods, represented by C . At this point, she would have maximized her utility (by being on the highest possible indifference curve U_2), given her budget constraint of \$20,000.

¹⁰ For a more extensive discussion of gasoline rationing, see H. E. Frech III and William C. Lee, "The Welfare Cost of Rationing-by-Queuing Across Markets: Theory and Estimates from the U.S. Gasoline Crises," *Quarterly Journal of Economics* (1987): 97–108.

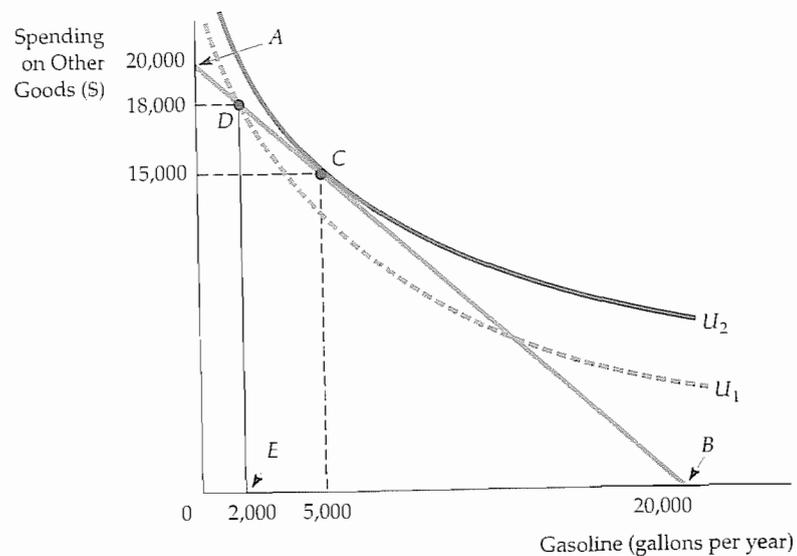


FIGURE 3.20 Inefficiency of Gasoline Rationing

When a good is rationed, less is available than consumers would like to buy. Consumers may be worse off. Without gasoline rationing, up to 20,000 gallons of gasoline are available for consumption (at point B). The consumer chooses point C on indifference curve U_2 , consuming 5,000 gallons of gasoline. However, with a limit of 2,000 gallons of gasoline under rationing (at point E), the consumer moves to D on the lower indifference curve U_1 .

With rationing, however, our consumer can purchase only 2,000 gallons of gasoline. Thus, she now faces budget line ADE, a line that is no longer a straight line because purchases above 2,000 gallons are not possible. The figure shows that her choice to consume at D involves a lower level of utility, U_1 , than would be achieved without rationing, U_2 , because she is consuming less gasoline and more of other goods than she would otherwise prefer.

*3.6 Cost-of-Living Indexes

The Social Security system has been the subject of heated debate for some time now. Under the present system, a retired person receives an annual benefit that is initially determined at the time of retirement and is based on his or her work history. The benefit then increases from year to year at a rate equal to the rate of increase of the Consumer Price Index (CPI). *The CPI is calculated each year by the U.S. Bureau of Labor Statistics as the ratio of the present cost of a typical bundle of consumer goods and services in comparison to the cost during a base period.* Does the CPI accurately reflect the cost of living for retirees? Is it appropriate to use the CPI as we now do—as a **cost-of-living index** for other government programs, for private union pensions, and for private wage agreements? The answers to these

questions lie in the economic theory of consumer behavior. In this section, we describe the theoretical underpinnings of cost indexes such as the CPI, using an example that describes the hypothetical price changes that students and their parents might face.

cost-of-living index Ratio of the present cost of a typical bundle of consumer goods and services compared with the cost during a base period.

Ideal Cost-of-Living Index

Let's look at two sisters, Rachel and Sarah, whose preferences are identical. When Sarah began her college education in 1990, her parents gave her a "discretionary" budget of \$500 per quarter. Sarah could spend the money on food, which was available at a price of \$2.00 per pound, and on books, which were available at a price of \$20 each. Sarah bought 100 pounds of food (at a cost of \$200) and 15 books (at a cost of \$300). Ten years later, in 2000 when Rachel (who had worked during the interim) is about to start college, her parents promise her a budget that is equivalent in buying power to that of her older sister. Unfortunately, prices in the college town have increased, with food now \$2.20 per pound and books \$100 each. By how much should the discretionary budget be increased to make Rachel as well off in 2000 as her sister Sarah was in 1990? Table 3.3 summarizes the relevant data and Figure 3.21 provides the answer.

The initial budget constraint facing Sarah in 1990 is given by line l_1 in Figure 3.21; her utility-maximizing combination of food and books is at point A on indifference curve U_1 . We can check that the cost of achieving this level of utility is \$500, as stated in the table:

$$\$500 = 100 \text{ lbs. of food} \times \$2.00/\text{lb.} + 15 \text{ books} \times \$20/\text{book}$$

As Figure 3.21 shows, for Rachel to achieve the same level of utility as Sarah while facing the new higher prices, she requires a budget sufficient to purchase the food-book consumption bundle given by point B on line l_2 (and tangent to indifference curve U_1), where she chooses 300 lbs. of food and 6 books. Note that in doing so, Rachel has taken into account the fact that the price of books has increased relative to food. Therefore she has substituted toward food and away from books.

The cost to Rachel of attaining the same level of utility as Sarah is given by

$$\$1,260 = 300 \text{ lbs. of food} \times \$2.20/\text{lb.} + 6 \text{ books} \times \$100/\text{book}$$

TABLE 3.3 Cost-of-Living Index		
	1990 (SARAH)	2000 (RACHEL)
Price of books	\$20/book	\$100/book
Number of books	15	6
Price of food	\$2.00/lb.	\$2.20/lb.
Pounds of food	100	300
Expenditure	\$500	\$1,260

In §1.1, we introduce the **Consumer Price Index** as a measure of the cost of a "typical" consumer's entire market basket. As such, changes in the CPI also measure the rate of inflation.

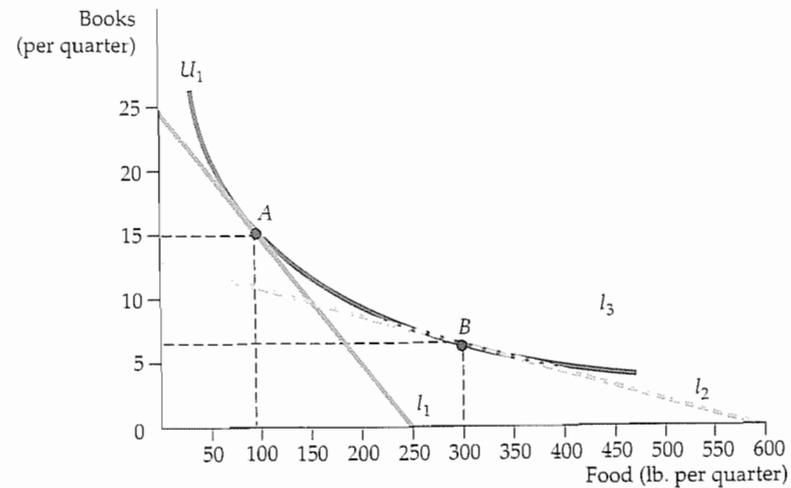


FIGURE 3.21 Cost-of-Living Indexes

The Laspeyres price index, which represents the cost of buying bundle A at current prices relative to the cost of bundle A at base-year prices, overstates the ideal cost-of-living index.

The ideal *cost-of-living adjustment* for Rachel is therefore \$760 (which is \$1,260 minus the \$500 that was given to Sarah). The ideal cost-of-living index is

$$\frac{\$1,260}{\$500} = 2.52$$

Like the CPI, our index needs a base year, which we will set at 1990 = 100, so that the value of the index in 2000 is 252. A value of 252 implies a 152 percent increase in the cost of living, whereas a value of 100 would imply that the cost of living has not changed. This **ideal cost-of-living index** represents the cost of attaining a given level of utility at current (2000) prices relative to the cost of attaining the same utility at base (1990) prices.

Laspeyres Index

Unfortunately, such an ideal cost-of-living index would entail large amounts of information. We would need to know individual preferences (which vary across the population) as well as prices and expenditures. Actual price indexes are therefore based on consumer purchases, not preferences. A price index, such as the CPI, which uses a *fixed consumption bundle in the base period*, is called a Laspeyres price index. The **Laspeyres price index** answers the question: *What is the amount of money at current year prices that an individual requires to purchase the bundle of goods and services that was chosen in the base year divided by the cost of purchasing the same bundle at base-year prices?*

Calculating a Laspeyres cost-of-living index for Rachel is a straightforward process. Buying 100 pounds of food and 15 books in 2000 would require an expenditure of \$1,720 ($100 \times \$2.20 + 15 \times \100). This expenditure allows Rachel to choose bundle A on budget line I_3 (or any other bundle on that line). Line I_3 was constructed by shifting line I_2 outward until it intersected point A. Note that I_3 is the budget line that allows Rachel to purchase, at current 2000

prices, the same consumption bundle that her sister purchased in 1990. To compensate Rachel for the increased cost of living, we must increase her discretionary budget by \$1,220. Using 100 as the base in 1990, the Laspeyres index is therefore

$$100 \times \$1,720/\$500 = 344$$

Comparing the Ideal Cost-of-Living and Laspeyres Indexes In our example, the Laspeyres price index is clearly much higher than the ideal price index. Does a Laspeyres index always overstate the true cost-of-living index? The answer is yes, as you can see from Figure 3.21. Suppose that Rachel was given the budget associated with line I_3 during the base year of 1990. She could choose bundle A, but clearly she could achieve a higher level of utility if she purchased more food and fewer books (by moving to the right on line I_3). Because A and B generate equal utility, it follows that Rachel is better off receiving a Laspeyres cost-of-living adjustment rather than an ideal adjustment. The Laspeyres index overcompensates Rachel for the higher cost of living, and the Laspeyres cost-of-living index is, therefore, greater than the ideal cost-of-living index. This result holds generally and applies to the CPI in particular. Why? Because the *Laspeyres price index assumes that consumers do not alter their consumption patterns as prices change*. By changing consumption, however—increasing purchases of items that have become relatively cheaper and decreasing purchases of relatively more expensive items—consumers can achieve the same level of utility without having to consume the same bundle of goods that they did before the price change.

Economic theory shows us that the Laspeyres cost-of-living index overstates the amount needed to compensate individuals for price increases. With respect to Social Security and other government programs, this means that *using the CPI to adjust retirement benefits will tend to overcompensate most recipients* and will thus require greater government expenditure. This is why the U.S. government has changed the construction of the CPI, switching from a Laspeyres price index to a more complex price index that reflects changing consumption patterns.

Paasche Index

Another commonly used cost-of-living index is the *Paasche index*. Unlike the Laspeyres index, which focuses on the cost of buying a base-year bundle, the **Paasche index** focuses on the cost of buying the *current year's bundle*. In particular, the Paasche index answers another question: *What is the amount of money at current year prices that an individual requires to purchase the current bundle of goods and services divided by the cost of purchasing the same bundle in the base year?*

Paasche index Amount of money at current-year prices that an individual requires to purchase a current bundle of goods and services divided by the cost of purchasing the same bundle in a base year.

Comparing the Laspeyres and Paasche Indexes It is helpful to compare the Laspeyres and the Paasche cost-of-living indexes.

■ **Laspeyres index:** The amount of money at current-year prices that an individual requires to purchase the bundle of goods and services that was chosen in the base year divided by the cost of purchasing the same bundle at base-year prices.

ideal cost-of-living index
Cost of attaining a given level of utility at current prices relative to the cost of attaining the same utility at base-year prices.

Laspeyres price index
Amount of money at current-year prices that an individual requires to purchase a bundle of goods and services chosen in a base year divided by the cost of purchasing the same bundle at base-year prices.

fixed-weight index Cost-of-living index in which the quantities of goods and services remain unchanged.

■ **Paasche index:** The amount of money at current-year prices that an individual requires to purchase the bundle of goods and services *chosen in the current year* divided by the cost of purchasing the same bundle in the base year.

Both the Laspeyres (LI) and Paasche (PI) indexes are **fixed-weight indexes**: The quantities of the various goods and services in each index remain unchanged. For the Laspeyres index, however, the quantities remain unchanged at *base-year* levels; for the Paasche they remain unchanged at *current-year* levels. Suppose generally that there are two goods, food (F) and clothing (C). Let:

P_{Ft} and P_{Ct} be current-year prices

P_{Fb} and P_{Cb} be base-year prices

F_t and C_t be current-year quantities

F_b and C_b be base-year quantities

We can write the two indexes as:

$$LI = \frac{P_{Ft}F_b + P_{Ct}C_b}{P_{Fb}F_b + P_{Cb}C_b}$$

$$PI = \frac{P_{Ft}F_t + P_{Ct}C_t}{P_{Fb}F_t + P_{Cb}C_t}$$

Just as the Laspeyres index will overstate the ideal cost of living, the Paasche will understate it because it assumes that the individual will buy the current year bundle in the base period. In actuality, facing base year prices, consumers would have been able to achieve the same level of utility at a lower cost by changing their consumption bundles. Because the Paasche index is a ratio of the cost of buying the current bundle divided by the cost of buying a base-year bundle, overstating the cost of the base-year bundle (the denominator in the ratio) will cause the index itself to be overstated.

To illustrate the Laspeyres-Paasche comparison, let's return to our earlier example and focus on Sarah's choices of books and food. For Sarah (who went to college in 1990), the cost of buying the base-year bundle of books and food at current-year prices is \$1,720 (100 lbs. \times \$2.20/lb. + 15 books \times \$100/book). The cost of buying the same bundle at base-year prices is \$500 (100 lbs \times \$2/lb. + 15 books \times \$20/book). The Laspeyres price index, LI, is therefore $100 \times \$1,720/\$500 = 344$, as reported previously. Likewise, the cost of buying the current-year bundle at current-year prices is \$1,260 (300 lbs. \times \$2.20/lb. + 6 books \times \$100/book). The cost of buying the same bundle at base-year prices is \$720 (300 lbs \times \$2/lb. + 6 books \times \$20/book). Consequently, the Paasche price index, PI, is $100 \times \$1,260/\$720 = 175$. As expected, the Paasche index is lower than the Laspeyres index.

Chain-Weighted Indexes

Neither the Laspeyres nor the Paasche index provides a perfect cost-of-living index, and the informational needs for the ideal index are too great. What is the best solution in practice? The U.S. government's most recent answer to this difficult question came in 1995, when it adopted the use of a **chain-weighted price index** to deflate its measure of gross domestic product (GDP) and thereby obtain an estimate of real GDP. Chain weighting was introduced to overcome problems

chain-weighted price index Cost-of-living index that accounts for changes in quantities of goods and services.

that arose when long-term comparisons of real GDP were made using fixed-weight price indexes (such as Paasche and Laspeyres) and prices were rapidly changing.

Economists have known for years that Laspeyres cost-of-living indexes overstate inflation. However, it was not until the energy price shocks of the 1970s, the more recent fluctuations in food prices, and the concern surrounding federal deficits that dissatisfaction with the Laspeyres index grew. It has been estimated, for example, that a failure to account for changes in computer-buying patterns in response to sharp decreases in computer prices has in recent years caused the CPI to overstate the cost of living substantially. As a result, the U.S. Bureau of Labor Statistics has been working to make improvements to the CPI.¹¹

EXAMPLE 3.7 The Bias in the CPI

In recent years, there has been growing public concern about the solvency of the Social Security system. At issue is the fact that retirement benefits are linked to the Consumer Price Index. Because the CPI is a Laspeyres index and can thus overstate the cost of living substantially, Congress has asked several economists to look into the matter.

A commission chaired by Stanford University professor Michael Boskin concluded that the CPI overstated inflation by approximately 1.1 percentage points—a significant amount given the relatively low rate of inflation in the United States in recent years.¹² According to the commission, approximately 0.4 percentage points of the 1.1-percentage-point bias was due to the failure of the Laspeyres price index to account for changes in the mix of consumption of the products in the base-year bundle. The remainder of the bias was due to the failure of the index to account for the growth of discount stores (approximately 0.1 percentage points), for improvements in the quality of existing products, and, most significantly, for the introduction of new products (0.6 percentage points).

If the bias in the CPI were to be eliminated, in whole or in part, the cost of a number of federal programs would decrease substantially (as would, of course, the corresponding benefits to eligible recipients in the programs). In addition to Social Security, affected programs include federal retirement programs (for railroad employees and military veterans), Supplemental Security Income (income support for the poor), food stamps, and child nutrition. According to one study, a 1-percentage-point reduction in the CPI would increase national savings and thereby reduce the national debt by approximately \$95 billion per year in year 2000 dollars.¹³

¹¹ Planned changes to the CPI are described by the Bureau of Labor Statistics in "Consumer Price Indexes: Overview of the 1998 revision of the Consumer Price Index," (at <http://stats.bls.gov/mlr/>) and in the Federal Reserve Bank of San Francisco Economic Letter No. 99-05 of February 5, 1999 (at <http://www.frbsf.org/econsrch/wklyltr/>).

¹² Michael J. Boskin, Ellen R. Dulberger, Robert J. Gordon, Zvi Griliches, and Dale W. Jorgenson, "The CPI Commission: Findings and Recommendations," *American Economic Review* 87, No. 2 (May 1997): 78-93.

¹³ Michael F. Bryan and Jagadeesh Gokhale, "The Consumer Price Index and National Savings," *Economic Commentary* (October 15, 1995) at <http://www.clev.frb.org/research/>. The data have been adjusted upward using the GDP deflator.

The effect of any CPI adjustments will not be restricted to the expenditure side of the federal budget. Because personal income tax brackets are inflation-adjusted, a CPI adjustment decreasing the rate of measured price increase would necessitate a smaller upper adjustment in tax brackets and, consequently, would increase federal tax revenues.

SUMMARY

1. The theory of consumer choice rests on the assumption that people behave rationally in an attempt to maximize the satisfaction that they can obtain by purchasing a particular combination of goods and services.
2. Consumer choice has two related parts: the study of the consumer's preferences and the analysis of the budget line that constrains the choices that a person can make.
3. Consumers make choices by comparing market baskets or bundles of commodities. Preferences are assumed to be complete (they can compare all possible market baskets) and transitive (if they prefer basket *A* to *B*, and *B* to *C*, then they prefer *A* to *C*). In addition, economists assume that more of each good is always preferred to less.
4. Indifference curves, which represent all combinations of goods and services that give the same level of satisfaction, are downward-sloping and cannot intersect one another.
5. Consumer preferences can be completely described by a set of indifference curves known as an indifference map. An indifference map provides an ordinal ranking of all choices that the consumer might make.
6. The marginal rate of substitution (MRS) of *F* for *C* is the maximum amount of *C* that a person is willing to give up to obtain 1 additional unit of *F*. The MRS diminishes as we move down along an indifference curve. When there is a diminishing MRS, preferences are convex.
7. Budget lines represent all combinations of goods for which consumers expend all their income. Budget lines shift outward in response to an increase in consumer income. When the price of one good (on the horizontal axis) changes while income and the price of the other good do not, budget lines pivot and rotate about a fixed point (on the vertical axis).
8. Consumers maximize satisfaction subject to budget constraints. When a consumer maximizes satisfaction by consuming some of each of two goods, the marginal rate of substitution is equal to the ratio of the prices of the two goods being purchased.
9. Maximization is sometimes achieved at a corner solution in which one good is not consumed. In such cases, the marginal rate of substitution need not equal the ratio of the prices.
10. The theory of revealed preference shows how the choices that individuals make when prices and income vary can be used to determine their preferences. When an individual chooses basket *A* even though she could afford *B*, we know that *A* is preferred to *B*.
11. The theory of the consumer can be presented by two different approaches. The indifference curve approach uses the ordinal properties of utility (that is, it allows for the ranking of alternatives). The utility function approach obtains a utility function by attaching a number to each market basket; if basket *A* is preferred to basket *B*, *A* generates more utility than *B*.
12. When risky choices are analyzed or when comparisons must be made among individuals, the cardinal properties of the utility function can be important. Usually the utility function will show diminishing marginal utility: As more and more of a good is consumed, the consumer obtains smaller and smaller increments of utility.
13. When the utility function approach is used and both goods are consumed, utility maximization occurs when the ratio of the marginal utilities of the two goods (which is the marginal rate of substitution) is equal to the ratio of the prices.
14. An ideal cost-of-living index measures the cost of buying, at current prices, a bundle of goods that generates the same level of utility as was provided by the bundle of goods consumed at base-year prices. The Laspeyres price index, however, represents the cost of buying the bundle of goods chosen in the base year at current prices relative to the cost of buying the same bundle at base-year prices. The CPI, like all Laspeyres price indexes, overstates the ideal cost-of-living index. By contrast, the Paasche index measures the cost at current-year prices of buying a bundle of goods chosen in the current year divided by the cost of buying the same bundle at base-year prices. It thus understates the ideal cost-of-living index.

QUESTIONS FOR REVIEW

1. What does *transitivity of preferences* mean?
2. Suppose that a set of indifference curves was not negatively sloped. What could you say about the desirability of the two goods?
3. Explain why two indifference curves cannot intersect.
4. Draw a set of indifference curves for which the marginal rate of substitution (MRS) is constant. Draw two budget lines with different slopes; show what the satisfaction-maximizing choice will be in each case. What conclusions can you draw?
5. Explain why a MRS between two goods must equal the ratio of the price of the goods for the consumer to achieve maximum satisfaction.
6. Explain why consumers are likely to be worse off when a product that they consume is rationed.
7. Upon merging with the West German economy, East German consumers indicated a preference for Mercedes-Benz automobiles over Volkswagens. However, when they converted their savings into deutsche marks, they flocked to Volkswagen dealerships. How can you explain this apparent paradox?
8. Describe the equal marginal principle. Explain why this principle may not hold if increasing marginal utility is associated with the consumption of one or both goods.
9. What is the difference between ordinal utility and cardinal utility? Explain why the assumption of cardinal utility is not needed in order to rank consumer choices.
10. The price of computers has fallen substantially over the past two decades. Use this drop in price to explain why the Consumer Price Index is likely to overstate substantially the cost-of-living index for individuals who use computers intensively.

EXERCISES

1. In this chapter, consumer preferences for various commodities did not change during the analysis. Yet in some situations, preferences do change as consumption occurs. Discuss why and how preferences might change over time with consumption of these two commodities:
 - a. cigarettes
 - b. dinner for the first time at a restaurant with a special cuisine.
2. Draw the indifference curves for the following individuals' preferences for two goods: hamburgers and beer.
 - a. Al likes beer but can live without hamburgers. He always prefers more beer no matter how many hamburgers he has.
 - b. Betty is indifferent between bundles of either three beers or two hamburgers. Her preferences do not change as she consumes any more of either food.
 - c. Chris eats one hamburger and washes it down with one beer. He will not consume an additional unit of one item without an additional unit of the other.
 - d. Doreen loves beer but is allergic to beef. Every time she eats a hamburger she breaks out in hives.
3. The price of tapes is \$10 and the price of CDs is \$15. Philip has a budget of \$100 and has already purchased 3 tapes. He thus has \$70 more to spend on additional tapes and CDs. Draw his budget line. If his remaining expenditure is made on 1 tape and 4 CDs, show Philip's consumption choice on the budget line.
4. Debra usually buys a soft drink when she goes to a movie theater, where she has a choice of three sizes. The 8-ounce drink costs \$1.50, the 12-ounce drink, \$2.00, and the 16-ounce drink, \$2.25. Describe the budget constraint that Debra faces when deciding how many ounces of the drink to purchase. (Assume Debra can costlessly dispose of any of the soft drink that she does not want.)
5. Suppose Bill views butter and margarine as perfectly substitutable for each other.
 - a. Draw a set of indifference curves that describes Bill's preferences for butter and margarine.
 - b. Are these indifference curves convex? Why?
 - c. If butter costs \$2 per package and margarine only \$1, and if Bill has a \$20 budget to spend for the month, which butter-margarine market basket will he choose? Can you show your answer graphically?
6. Suppose Jones and Smith have decided to allocate \$1,000 per year to liquid refreshment in the form of alcoholic or nonalcoholic drinks. Jones and Smith differ substantially in their preferences for these two forms of refreshment. Jones prefers alcoholic to nonalcoholic drinks, while Smith prefers the nonalcoholic option.
 - a. Draw a set of indifference curves for Jones and a second set for Smith.
 - b. Using the concept of marginal rate of substitution, explain why the two sets of curves are different from each other.

- c. If both Smith and Jones pay the same prices for their refreshments, will their marginal rates of substitution of alcoholic for nonalcoholic drinks be the same or different? Explain.
7. Consumers in Georgia pay twice as much for avocados as they do for peaches. However, avocados and peaches are equally priced in California. If consumers in both states maximize utility, will the marginal rates of substitution of peaches for avocados be the same for consumers in both states? If not, which will be higher?
8. Anne is a frequent flyer whose fares are reduced (through coupon giveaways) by 25 percent after she flies 25,000 miles a year and then by 50 percent after she flies 50,000 miles. Can you graph the budget line that Anne faces in making her flight plans for the year?
9. Antonio buys 8 new college textbooks during his first year at school at a cost of \$50 each. Used books cost only \$30 each. When the bookstore announces that there will be a 20-percent price increase in new texts and a 10-percent increase in used texts for the coming year, Antonio's father offers him \$80 extra. Is Antonio better off or worse off after the price change?
10. Suppose that Samantha and Jason both spend \$24 per week on video and movie entertainment. When the prices of videos and movies are both \$4, they each rent 3 videos and buy 3 movie tickets. Following a video price war and an increase in the cost of movie tickets, the price of videos falls to \$2 while the price of movie tickets increases to \$6. Samantha now rents 6 videos and buys 2 movie tickets; Jason, however, buys 1 movie ticket and rents 9 videos.
- a. Is Samantha better off or worse off after the price change?
- b. Is Jason better off or worse off?
11. Connie allocates \$200 of her monthly food budget between two goods: meat and potatoes.
- a. Suppose meat costs \$4 per pound and potatoes \$2 per pound. Draw Connie's budget constraint.
- b. Suppose also that her utility function is given by the equation $u(M,P) = 2M + P$. What combination of meat and potatoes should she buy to maximize her utility? (*Hint:* Meat and potatoes are perfect substitutes.)
- c. Connie's supermarket is running a special promotion: If she buys 20 pounds of potatoes (at \$2 per pound), she gets the next 10 pounds for free. This offer applies only to the first 20 pounds she buys. All potatoes in excess of the first 20 pounds (excluding bonus potatoes) are still \$2 per pound. Draw her budget constraint.
- d. When an outbreak of potato rot raises the price of potatoes to \$4 per pound, the supermarket ends its promotion. What does Connie's budget constraint look like now? What combination of meat and potatoes will maximize her utility?
12. The utility that Jane receives by consuming food F and clothing C is given by $u(F,C) = FC$.
- a. Draw the indifference curve associated with a utility level of 12 and the indifference curve associated with a utility level of 24. Are the indifference curves convex?
- b. Suppose that food costs \$1 a unit and clothing \$3 a unit. Jane has \$12 to spend on food and clothing. Graph the budget line that she faces.
- c. What is the utility-maximizing choice of food and clothing? (*Hint:* Solve the problem graphically.)
- d. What is the marginal rate of substitution of food for clothing when utility is maximized?
- e. Suppose that Jane buys 3 units of food and 3 units of clothing with her \$12 budget. Would her marginal rate of substitution of food for clothing be greater or less than $1/3$? Explain.
13. The utility that Meredith receives by consuming food F and clothing C is given by $u(F,C) = FC$. Suppose that her income in 1990 is \$1,200 and that the prices of food and clothing are \$1 per unit of each. By the year 2000, however, the price of food has increased to \$2 and clothing to \$3. Let 100 represent the cost-of-living index for 1990. Calculate both the ideal and the Laspeyres cost-of-living index for Meredith for 2000. (*Hint:* Meredith will spend equal amounts on food and clothing.)

CHAPTER 4

Individual and Market Demand

Chapter 3 laid the foundation for the theory of consumer demand. We discussed the nature of consumers' preferences and saw how, given budget constraints, consumers choose market baskets that maximize utility. From here it's a short step to analyzing demand itself and showing how the demand for a good depends on its price, the prices of other goods, and income.

Our analysis of demand proceeds in six steps:

1. We begin by deriving the demand curve for an individual consumer. Because we know how changes in price and income affect a person's budget line, we can determine how they affect consumption choice. We will use this information to see how the quantity of a good demanded varies in response to price changes as we move along an individual's demand curve. We will also see how this demand curve shifts in response to changes in the individual's income.
2. With this foundation, we will examine the effect of a price change in more detail. When the price of a good goes up, individual demand for it can change in two ways. First, because it has now become more expensive relative to other goods, consumers will buy less of it and more of other goods. Second, the higher price reduces the consumer's purchasing power. This reduction is just like a reduction in income and will lead to a reduction in the consumer's demand. By analyzing these two distinct effects, we will better understand the characteristics of demand.
3. Next, we will see how individual demand curves can be aggregated to determine the market demand curve. We will also study the characteristics of market demand and see why the demands for some kinds of goods differ considerably from the demands for others.
4. We will go on to show how market demand curves can be used to measure the benefits that people receive when they consume products, above and beyond the expenditures they make. This information will be especially important later, when we study the effects of government intervention in a market.
5. We then describe the effects of *network externalities*—i.e., what happens when a person's demand for a good also

Chapter Outline

- 4.1 Individual Demand 102
 - 4.2 Income and Substitution Effects 110
 - 4.3 Market Demand 116
 - 4.4 Consumer Surplus 123
 - 4.5 Network Externalities 127
 - *4.6 Empirical Estimation of Demand 131
- Appendix: Demand Theory—A Mathematical Treatment 139

List of Examples

- 4.1 Consumer Expenditures in the United States 108
- 4.2 The Effects of a Gasoline Tax 114
- 4.3 The Aggregate Demand for Wheat 120
- 4.4 The Demand for Housing 122
- 4.5 The Value of Clean Air 125
- 4.6 Network Externalities and the Demands for Computers and E-Mail 130
- 4.7 The Demand for Ready-to-Eat Cereal 134

depends on the demands of *other* people. These effects play a crucial role in the demands for many high-tech products, such as computer hardware and software, and telecommunications systems.

- Finally, we will briefly describe some of the methods that economists use to obtain empirical information about demand.

4.1 Individual Demand

This section shows how the demand curve of an individual consumer follows from the consumption choices that a person makes when faced with a budget constraint. To illustrate these concepts graphically, we will limit the available goods to food and clothing and will rely on the utility-maximization approach described in Section 3.3.

Price Changes

We begin by examining ways in which the consumption of food and clothing changes when the price of food changes. Figure 4.1 shows the consumption choices that a person will make when allocating a fixed amount of income between the two goods.

Initially, the price of food is \$1, the price of clothing \$2, and the consumer's income \$20. The utility-maximizing consumption choice is at point *B* in Figure 4.1(a). Here, the consumer buys 12 units of food and 4 units of clothing, thus achieving the level of utility associated with indifference curve U_2 .

Now look at Figure 4.1(b), which shows the relationship between the price of food and the quantity demanded. The horizontal axis measures the quantity of food consumed, as in Figure 4.1(a), but the vertical axis now measures the price of food. Point *G* in Figure 4.1(b) corresponds to point *B* in Figure 4.1(a). At *G*, the price of food is \$1, and the consumer purchases 12 units of food.

Suppose the price of food increases to \$2. As we saw in Chapter 3, the budget line in Figure 4.1(a) rotates inward about the vertical intercept, becoming twice as steep as before. The higher relative price of food has increased the magnitude of the slope of the budget line. The consumer now achieves maximum utility at *A*, which is found on a lower indifference curve, U_1 . (Because the price of food has risen, the consumer's purchasing power—and thus attainable utility—has fallen.) At *A*, the consumer chooses 4 units of food and 6 units of clothing. In Figure 4.1(b), this modified consumption choice is at *E*, which shows that at a price of \$2, 4 units of food are demanded.

Finally, what will happen if the price of food *decreases* to 50 cents? Because the budget line now rotates outward, the consumer can achieve the higher level of utility associated with indifference curve U_3 in Figure 4.1(a) by selecting *D*, with 20 units of food and 5 units of clothing. Point *H* in Figure 4.1(b) shows the price of 50 cents and the quantity demanded of 20 units of food.

The Individual Demand Curve

We can go on to include all possible changes in the price of food. In Figure 4.1(a), the **price-consumption curve** traces the utility-maximizing combinations of food and clothing associated with every possible price of food. Note that as the price of food falls, attainable utility increases and the consumer buys more food.

In §3.3, we explain how consumers choose the market basket on the highest indifference curve that touches the consumer's budget line.

In §3.2, we explain how the budget line shifts in response to a price change.

price-consumption curve
Curve tracing the utility-maximizing combinations of two goods as the price of one changes.

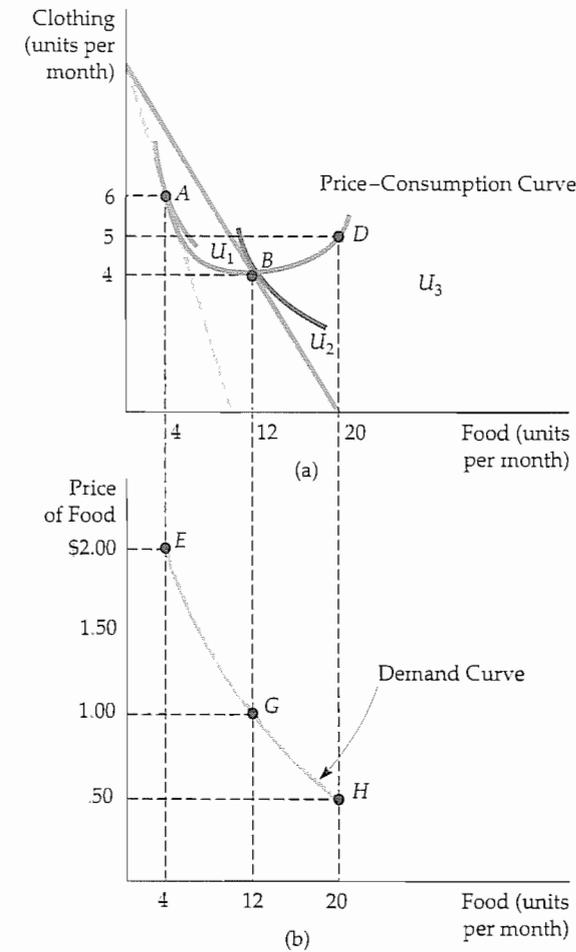


FIGURE 4.1 Effect of Price Changes

A reduction in the price of food, with income and the price of clothing fixed, causes this consumer to choose a different market basket. In (a), the baskets that maximize utility for various prices of food (point *A*, \$2; *B*, \$1; *D*, \$0.50) trace out the price-consumption curve. Part (b) gives the demand curve, which relates the price of food to the quantity demanded. (Points *E*, *G*, and *H* correspond to points *A*, *B*, and *D*, respectively.)

This pattern of increasing consumption of a good in response to a decrease in price almost always holds. But what happens to the consumption of clothing as the price of food falls? As Figure 4.1(a) shows, the consumption of clothing may either increase or decrease. Both food *and* clothing consumption can increase because the decrease in the price of food has increased the consumer's ability to purchase both goods.

An **individual demand curve** relates the quantity of a good that a single consumer will buy to the price of that good. In Figure 4.1(b), the individual demand curve relates the quantity of food that the consumer will buy to the price of food. This demand curve has two important properties.

individual demand curve
Curve relating the quantity of a good that a single consumer will buy to its price.

1. *The level of utility that can be attained changes as we move along the curve.*
The lower the price of the product, the higher its level of utility. Note from Figure 4.1(a) that a higher indifference curve is reached as the price falls. Again, this result simply reflects the fact that as the price of a product falls, the consumer's purchasing power increases.
2. *At every point on the demand curve, the consumer is maximizing utility by satisfying the condition that the marginal rate of substitution (MRS) of food for clothing equals the ratio of the prices of food and clothing.* As the price of food falls, the price ratio and the MRS also fall. In Figure 4.1, the price ratio falls from 1 (\$2/\$2) at E (because the curve U_1 is tangent to a budget line with a slope of -1 at A) to $1/2$ (\$1/\$2) at G, to $1/4$ (\$0.50/\$2) at H. Because the consumer is maximizing utility, the MRS of food for clothing decreases as we move down the demand curve. This phenomenon makes intuitive sense because it tells us that the relative value of food falls as the consumer buys more of it.

In §3.1, we introduce the marginal rate of substitution as a measure of the maximum amount of one good that the consumer is willing to give up in order to obtain one unit of another good.

The fact that the MRS varies along the individual's demand curve tells us something about how consumers value the consumption of a good or service. Suppose we were to ask a consumer how much she would be willing to pay for an additional unit of food when she is currently consuming 4 units. Point E on the demand curve in Figure 4.1(b) provides the answer: \$2. Why? As we pointed out above, because the MRS of food for clothing is 1 at E, one additional unit of food is worth one additional unit of clothing. But a unit of clothing costs \$2, which is, therefore, the value (or marginal benefit) obtained by consuming an additional unit of food. Thus, as we move down the demand curve in Figure 4.1(b), the MRS falls. Likewise, the value that the consumer places on an additional unit of food falls from \$2 to \$1 to \$0.50.

Income Changes

We have seen what happens to the consumption of food and clothing when the price of food changes. Now let's see what happens when income changes.

The effects of a change in income can be analyzed in much the same way as a price change. Figure 4.2(a) shows the consumption choices that a consumer will make when allocating a fixed income to food and clothing when the price of food is \$1 and the price of clothing \$2. As in Figure 4.1(a), the quantity of clothing is measured on the vertical axis and the quantity of food on the horizontal axis. Income changes appear as changes in the budget line. Initially, the consumer's income is \$10. The utility-maximizing consumption choice is then at A, at which she buys 4 units of food and 3 units of clothing.

This choice of 4 units of food is also shown in Figure 4.2(b) as E on demand curve D_1 . Demand curve D_1 is the curve that would be traced out if we held income fixed at \$10 but varied the price of food. Because we are holding the price of food constant, we will observe only a single point E on this demand curve.

What happens if the consumer's income is increased to \$20? Her budget line then shifts outward parallel to the original budget line, allowing her to attain the utility level associated with indifference curve U_2 . Her optimal consumption choice is now at B, where she buys 10 units of food and 5 units of clothing. In Figure 4.2(b) her consumption of food is shown as G on demand curve D_2 . D_2 is the demand curve that would be traced out if we held income fixed at \$20 but varied the price of food. Finally, note that if her income increases to \$30, she chooses D, with a market basket containing 16 units of food (and 7 units of clothing), represented by H in Figure 4.2(b).

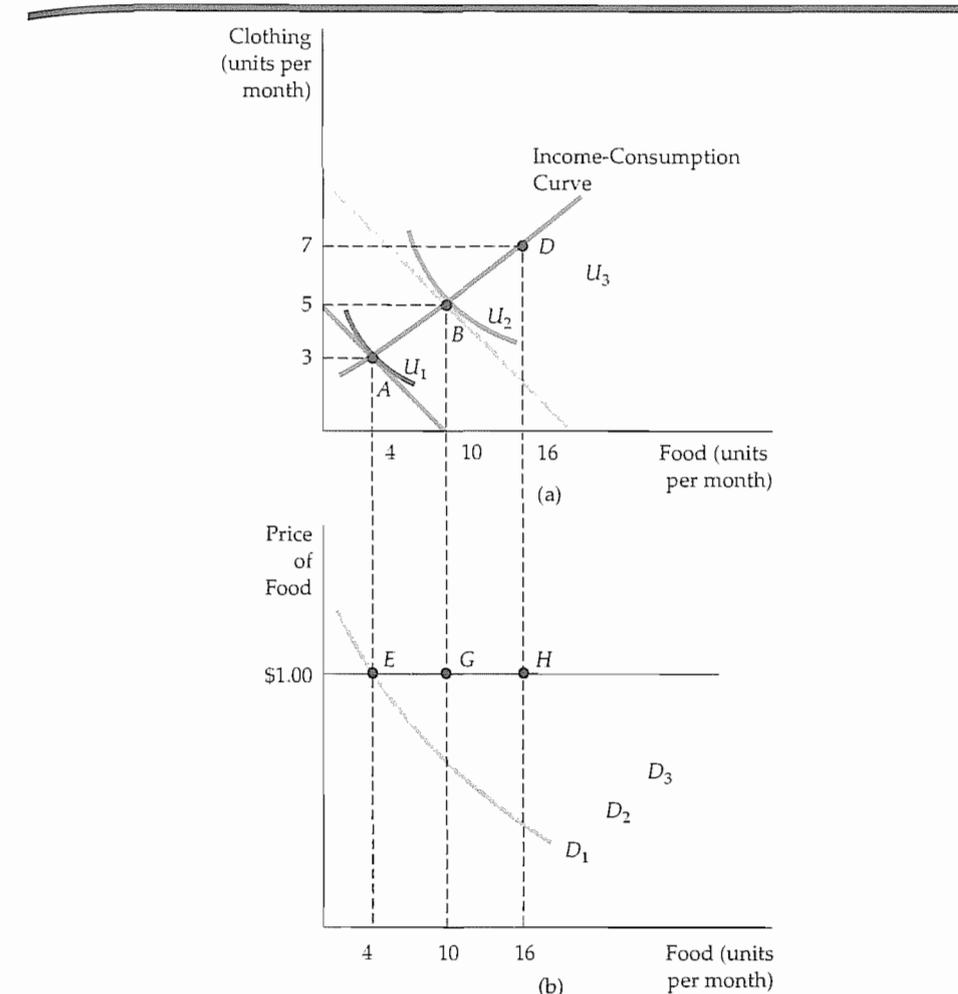


FIGURE 4.2 Effect of Income Changes

An increase in income, with the prices of all goods fixed, causes consumers to alter their choices of market basket. In part (a), the baskets that maximize consumer satisfaction for various incomes (point A, \$10; B, \$20; D, \$30) trace out the income-consumption curve. The shift to the right of the demand curve in response to the increases in income is shown in part (b). (Points E, G, and H correspond to points A, B, and D, respectively.)

We could go on to include all possible changes in income. In Figure 4.2(a), the **income-consumption curve** traces out the utility-maximizing combinations of food and clothing associated with every income level. The income-consumption curve in Figure 4.2 slopes upward because the consumption of both food and clothing increases as income increases. Previously, we saw that a change in the price of a good corresponds to a *movement along a demand curve*. Here, the situation is different. Because each demand curve is measured for a particular level of income, any change in income must lead to a *shift in the demand curve itself*. Thus A on the income-consumption curve in Figure 4.2(a) corresponds to E on demand curve D_1 in Figure 4.2(b); B corresponds to G on a different demand

income-consumption curve
Curve tracing the utility-maximizing combinations of two goods as a consumer's income changes.

curve D_2 . The upward-sloping income-consumption curve implies that an increase in income causes a shift to the right in the demand curve—in this case from D_1 to D_2 to D_3 .

Normal versus Inferior Goods

When the income-consumption curve has a positive slope, the quantity demanded increases with income. As a result, the income elasticity of demand is positive. The greater the shifts to the right of the demand curve, the larger the income elasticity. In this case, the goods are described as *normal*: Consumers want to buy more of them as their income increases.

In some cases, the quantity demanded *falls* as income increases; the income elasticity of demand is negative. We then describe the good as *inferior*. The term *inferior* simply means that consumption falls when income rises. Hamburger, for example, is inferior for some people: As their income increases, they buy less hamburger and more steak.

Figure 4.3 shows the income-consumption curve for an inferior good. For relatively low levels of income, both hamburger and steak are normal goods. As income rises, however, the income-consumption curve bends backward (from point B to C). This shift occurs because hamburger has become an inferior good—its consumption has fallen as income has increased.

Engel Curves

Income-consumption curves can be used to construct **Engel curves**, which relate the quantity of a good consumed to an individual's income. Figure 4.4 shows

In §2.3, we explain that the income elasticity of demand is the percentage change in the quantity demanded resulting from a 1-percent increase in income.

Engel curve Curve relating the quantity of a good consumed to income.

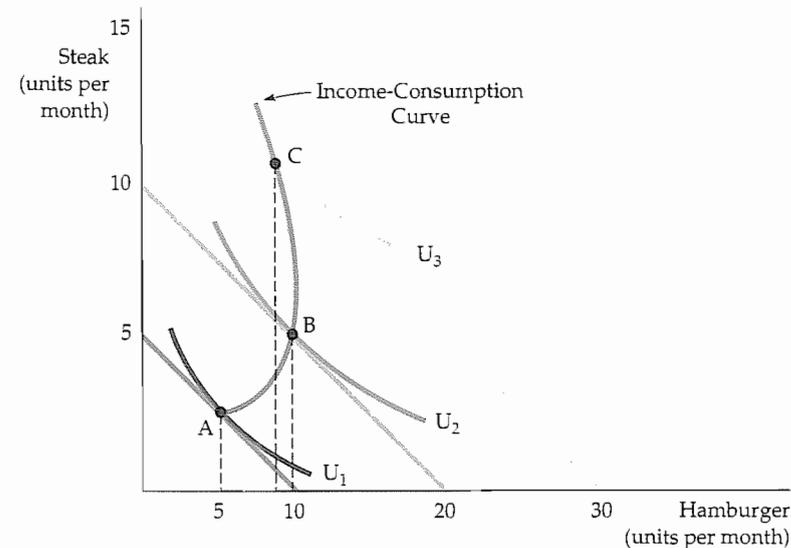
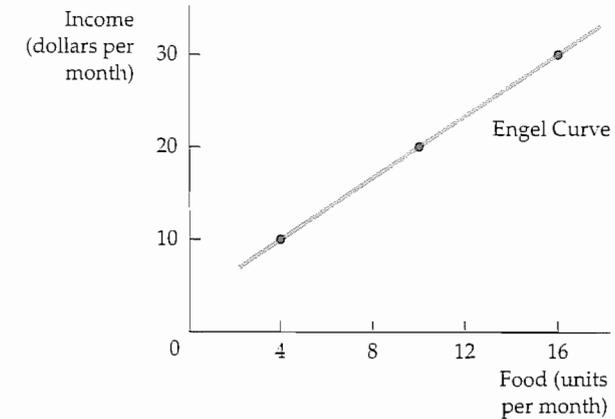
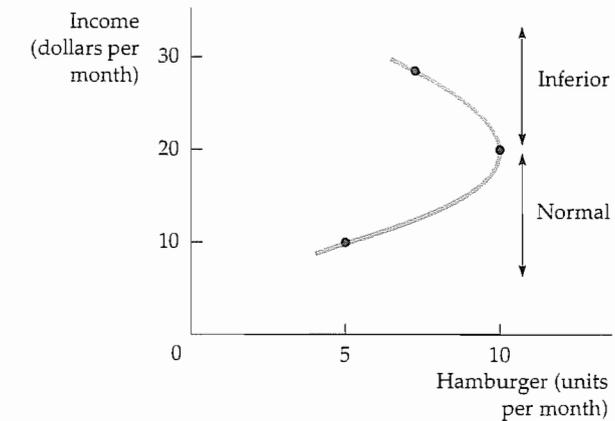


FIGURE 4.3 An Inferior Good

An increase in a person's income can lead to less consumption of one of the two goods being purchased. Here, hamburger, though a normal good between A and B , becomes an inferior good when the income-consumption curve bends backward between B and C .



(a)



(b)

FIGURE 4.4 Engel Curves

Engel curves relate the quantity of a good consumed to income. In (a), food is a normal good and the Engel curve is upward sloping. In (b), however, hamburger is a normal good for income less than \$20 per month and an inferior good for income greater than \$20 per month.

how such curves are constructed for two different goods. Figure 4.4(a), which shows an upward-sloping Engel curve, is derived directly from Figure 4.2(a). In both figures, as the individual's income increases from \$10 to \$20 to \$30, her consumption of food increases from 4 to 10 to 16 units. Recall that in Figure 4.2(a) the vertical axis measured units of clothing consumed per month and the horizontal axis units of food per month; changes in income were reflected as shifts in the budget line. In Figures 4.4(a) and (b), we have replotted the data to put income on the vertical axis while keeping food and hamburger on the horizontal.

The upward-sloping Engel curve in Figure 4.4(a)—like the upward-sloping income-consumption curve in Figure 4.2(a)—applies to all normal goods. Note that an Engel curve for clothing would have a similar shape (clothing consumption increases from 3 to 5 to 7 units as income increases).

Figure 4.4(b), derived from Figure 4.3, shows the Engel curve for hamburger. We see that hamburger consumption increases from 5 to 10 units as income increases from \$10 to \$20. As income increases further, from \$20 to \$30, consumption falls to 8 units. The portion of the Engel curve that slopes downward is the income range in which hamburger is an inferior good.

EXAMPLE 4.1 Consumer Expenditures in the United States

The Engel curves we just examined apply to individual consumers. However, we can also derive Engel curves for groups of consumers. This information is particularly useful if we want to see how consumer spending varies among different income groups. Table 4.1 illustrates these spending patterns for several items taken from a survey by the U.S. Bureau of Labor Statistics. Although the data are averaged over many households, they can be interpreted as describing the expenditures of a typical family.

Note that the data relate *expenditures* on a particular item rather than the *quantity* of the item to income. The first two items, entertainment and owned dwellings, are consumption goods for which the income elasticity of demand is high. Average family expenditures on entertainment increase almost eightfold when we move from the lowest to highest income group. The same pattern applies to the purchase of homes: There is a more than tenfold increase in expenditures from the lowest to the highest category.

In contrast, expenditures on *rental* housing actually *fall* with income. This pattern reflects the fact that most higher-income individuals own rather than rent homes. Thus rental housing is an inferior good, at least for incomes above \$30,000 per year. Finally, note that health care, food, and clothing are consumption items for which the income elasticities are positive, but not as high as for entertainment or owner-occupied housing.

The data in Table 4.1 have been plotted in Figure 4.5 for rented dwellings, health care, and entertainment. Observe in the three Engel curves that as

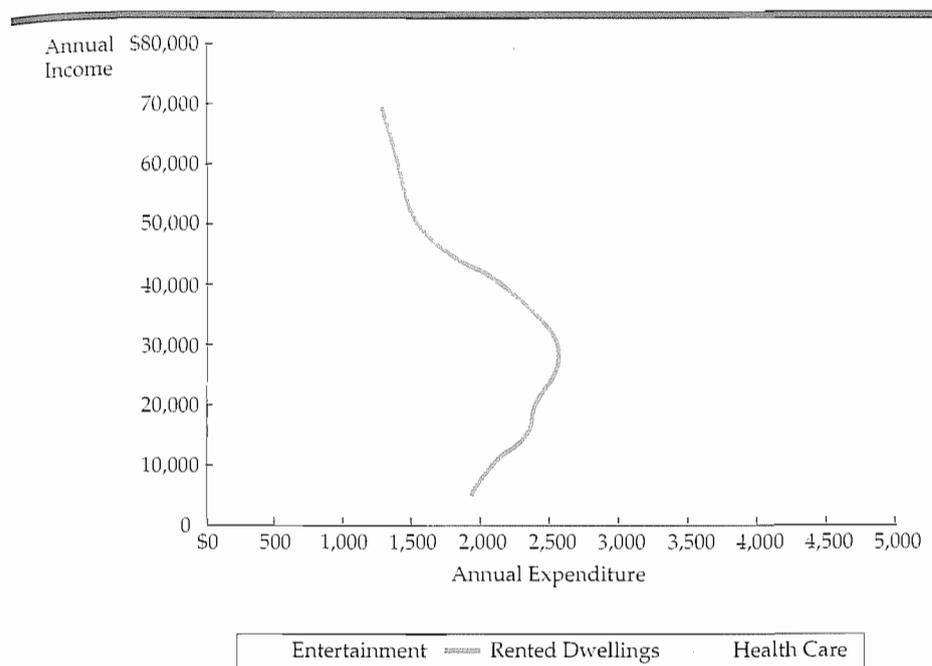


FIGURE 4.5 Engel Curves for U.S. Consumers

Average per-capita expenditures on rented dwellings, health care, and entertainment are plotted as functions of annual income. Health care and entertainment are superior goods: Expenditures increase with income. Rental housing, however, is an inferior good for incomes above \$30,000.

income rises, expenditures on entertainment increase rapidly while expenditures on rental housing increase when income is low, but decrease once income exceeds \$30,000.

EXPENDITURES (\$) ON:	INCOME GROUP (1997 \$)						
	LESS THAN 10,000	10,000–19,000	20,000–29,000	30,000–39,000	40,000–49,000	50,000–69,000	70,000 AND ABOVE
Entertainment	700	947	1,274	1,514	2,054	2,654	4,300
Owned dwellings	1,116	1,725	2,253	3,243	4,454	5,793	9,898
Rented dwellings	1,957	2,170	2,371	2,536	2,137	1,540	1,266
Health care	1,031	1,697	1,918	1,820	2,052	2,214	2,642
Food	2,656	3,385	4,109	4,888	5,429	6,220	8,279
Clothing	859	978	1,363	1,772	1,778	2,614	3,442

Source: U.S. Department of Labor, Bureau of Labor Statistics, "Consumer Expenditure Survey: 1997."

Substitutes and Complements

The demand curves that we graphed in Chapter 2 showed the relationship between the price of a good and the quantity demanded, with preferences, income, and the prices of all other goods held constant. For many goods, demand is related to the consumption and prices of other goods. Baseball bats and baseballs, hot dogs and mustard, and computer hardware and software are all examples of goods that tend to be used together. Other goods, such as cola and diet cola, owner-occupied houses and rental apartments, movie tickets and videocassette rentals, tend to substitute for one another.

Recall from Section 2.4 that two goods are *substitutes* if an increase in the price of one leads to an increase in the quantity demanded of the other. If the price of a movie ticket rises, we would expect individuals to rent more videos, because movie tickets and videos are substitutes. Similarly, two goods are *complements* if an increase in the price of one good leads to a decrease in the quantity demanded of the other. If the price of gasoline goes up, causing gasoline consumption to fall, we would expect the consumption of motor oil to fall as well, because gasoline and motor oil are used together. Two goods are *independent* if a change in the price of one good has no effect on the quantity demanded of the other.

One way to see whether two goods are complements or substitutes is to examine the price-consumption curve. Look again at Figure 4.1. Note that in the downward-sloping portion of the price-consumption curve, food and clothing are substitutes: The lower price of food leads to a lower consumption of clothing (perhaps because as food expenditures increase, less income is available to spend on clothing). Similarly, food and clothing are complements in the upward-sloping portion of the curve: The lower price of food leads to higher clothing consumption (perhaps because the consumer eats more meals at restaurants and must be suitably dressed).

The fact that goods can be complements or substitutes suggests that when studying the effects of price changes in one market, it may be important to look at the consequences in related markets. (Interrelationships among markets are discussed in more detail in Chapter 16.) Determining whether two goods are complements, substitutes, or independent goods is ultimately an empirical question. To answer the question, we need to look at the ways in which the demand for the first good shifts (if at all) in response to a change in the price of the second. This question is more difficult than it sounds because lots of things are likely to be changing at the same time that the price of the first good changes. In fact, Section 6 of this chapter is devoted to examining ways to distinguish empirically among the many possible explanations for a change in the demand for the second good. First, however, it will be useful to undertake a basic theoretical exercise. In the next section, we delve into the ways in which a change in the price of a good can affect consumer demand.

4.2 Income and Substitution Effects

A fall in the price of a good has two effects:

1. Consumers will tend to buy more of the good that has become cheaper and less of those goods that are now relatively more expensive. This response to the change in the relative prices of goods is called the *substitution effect*.
2. Because one of the goods is now cheaper, consumers enjoy an increase in real purchasing power. They are better off because they can buy the same amount of the good for less money and thus have money left over for additional purchases. The change in demand resulting from this change in real purchasing power is called the *income effect*.

Normally, these two effects occur simultaneously, but it will be useful to distinguish between them for purposes of analysis. The specifics are illustrated in Figure 4.6, where the initial budget line is RS and there are two goods, food and clothing. Here, the consumer maximizes utility by choosing the market basket at A , thereby obtaining the level of utility associated with the indifference curve U_1 .

Now, let's see what happens if the *price of food falls*, causing the budget line to rotate outward to line RT . The consumer now chooses the market basket at B on indifference curve U_2 . Because market basket B was chosen even though market basket A was feasible, we know (from our discussion of revealed preference in Section 3.4) that B is preferred to A . Thus the reduction in the price of food allows the consumer to increase her level of satisfaction—her purchasing power has increased. The total change in the consumption of food caused by the lower

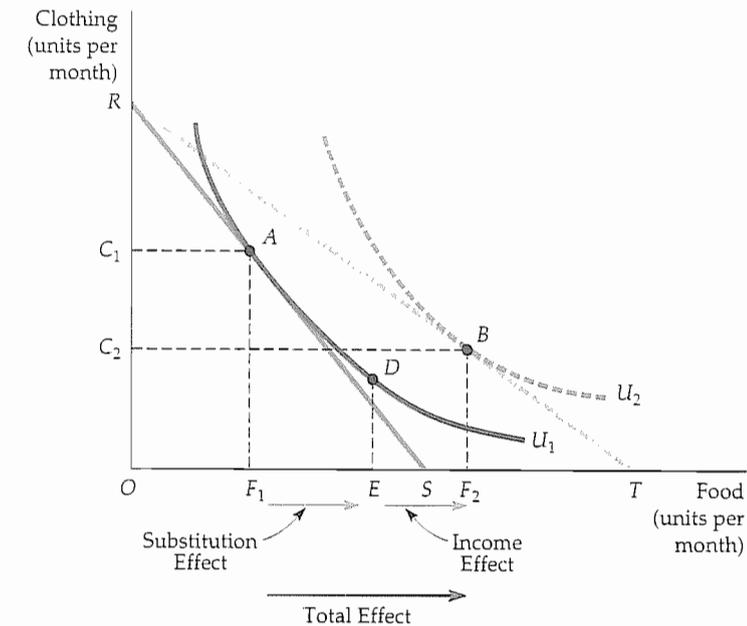


FIGURE 4.6 Income and Substitution Effects: Normal Good

A decrease in the price of food has an income effect and a substitution effect. The consumer is initially at A on budget line RS . When the price of food falls, consumption increases by F_1F_2 as the consumer moves to B . The substitution effect F_1E (associated with a move from A to D) changes the relative prices of food and clothing but keeps real income (satisfaction) constant. The income effect EF_2 (associated with a move from D to B) keeps relative prices constant but increases purchasing power. Food is a normal good because the income effect EF_2 is positive.

price is given by F_1F_2 . Initially, the consumer purchased OF_1 units of food, but after the price change, food consumption has increased to OF_2 . Line segment F_1F_2 , therefore, represents the increase in desired food purchases.

Substitution Effect

The drop in price has both a substitution effect and an income effect. The **substitution effect** is the change in food consumption associated with a change in the price of food, with the level of utility held constant. The substitution effect captures the change in food consumption that occurs as a result of the price change that makes food relatively cheaper than clothing. This substitution is marked by a movement along an indifference curve. In Figure 4.6, the substitution effect can be obtained by drawing a budget line which is parallel to the new budget line RT (reflecting the lower relative price of food) but which is just tangent to the original indifference curve U_1 (holding the level of satisfaction constant). The new, lower imaginary budget line reflects the fact that nominal income was reduced in order to accomplish our conceptual goal of isolating the substitution effect. Given that budget line, the consumer chooses market basket D and consumes OE units of food. The line segment F_1E thus represents the substitution effect.

Figure 4.6 makes it clear that when the price of food declines, the substitution effect always leads to an increase in the quantity of food demanded. The explanation lies in our fourth assumption about consumer preferences in Section

substitution effect Change in consumption of a good associated with a change in its price, with the level of utility held constant.

In §3.4, we show how information about consumer preferences is revealed by consumption choices made.

3.1—namely, that preferences are convex. Thus, with the convex indifference curves shown in the figure, the point that maximizes satisfaction on the new budget line RT must lie below and to the right of the original point of tangency.

Income Effect

income effect Change in consumption of a good resulting from an increase in purchasing power, with relative price held constant.

Now consider the **income effect**: the change in food consumption brought about by the increase in purchasing power, with the price of food held constant. In Figure 4.6, the income effect can be seen by moving from the imaginary budget line that passes through point D to the original budget line, RT , that passes through B . The consumer chooses market basket B on indifference curve U_2 (because the lower price of food has increased her level of utility). The increase in food consumption from OE to OF_2 is the measure of the income effect, which is positive, because food is a *normal good* (consumers will buy more of it as their incomes increase). Because it reflects a movement from one indifference curve to another, the income effect measures the change in the consumer's purchasing power.

We have seen that the total effect of a change in price is given theoretically by the sum of the substitution effect and the income effect:

$$\text{Total Effect } (F_1F_2) = \text{Substitution Effect } (F_1E) + \text{Income Effect } (EF_2)$$

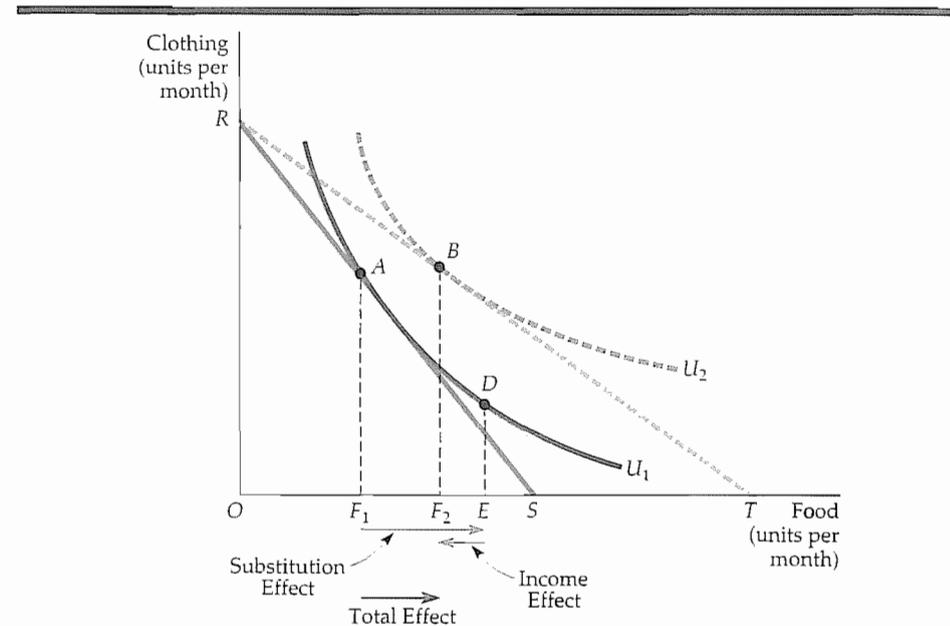


FIGURE 4.7 Income and Substitution Effects: Inferior Good

The consumer is initially at A on budget line RS . With a decrease in the price of food, the consumer moves to B . The resulting change in food purchased can be broken down into a substitution effect F_1E (associated with a move from A to D) and an income effect EF_2 (associated with a move from D to B). In this case, food is an inferior good because the income effect is negative. However, because the substitution effect exceeds the income effect, the decrease in the price of food leads to an increase in the quantity of food demanded.

Recall that the direction of the substitution effect is always the same: A decline in price leads to an increase in consumption of the good. However, the income effect can move demand in either direction, depending on whether the good is normal or inferior.

A good is *inferior* when the income effect is negative: As income rises, consumption falls. Figure 4.7 shows income and substitution effects for an inferior good. The negative income effect is measured by line segment EF_2 . Even with inferior goods, the income effect is rarely large enough to outweigh the substitution effect. As a result, when the price of an inferior good falls, its consumption almost always increases.

A Special Case: The Giffen Good

Theoretically, the income effect may be large enough to cause the demand curve for a good to slope upward. We call such a good a **Giffen good**, and Figure 4.8 shows its income and substitution effects. Initially, the consumer is at A , consuming relatively little clothing and much food. Now the price of food declines. The decline in the price of food frees enough income so that the consumer desires to buy more clothing and fewer units of food, as illustrated by B . Revealed preference tells us that the consumer is better off at B rather than A even though less food is consumed.

Giffen good Good whose demand curve slopes upward because the (positive) income effect is larger than the (negative) substitution effect.

Though intriguing, the Giffen good is rarely of practical interest because it requires a large negative income effect. But the income effect is usually small: Individually, most goods account for only a small part of a consumer's budget. Large income effects are often associated with normal rather than inferior goods (e.g., total spending on food or housing).

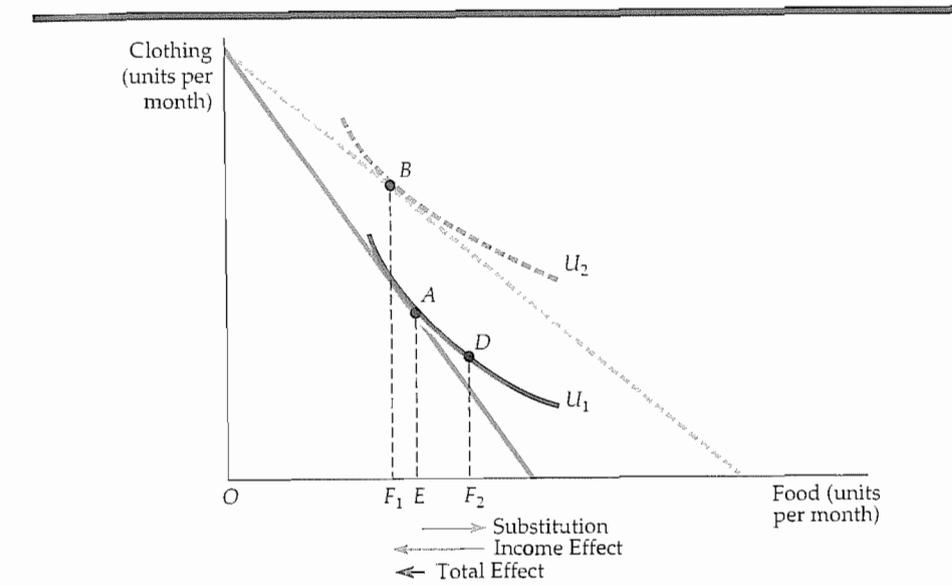


FIGURE 4.8 Upward-Sloping Demand Curve: The Giffen Good

When food is an inferior good, and when the income effect is large enough to dominate the substitution effect, the demand curve will be upward-sloping. The consumer is initially at point A but, after the price of food falls, moves to B and consumes less food. Because the income effect F_2F_1 is larger than the substitution effect EF_2 , the decrease in the price of food leads to a lower quantity of food demanded.

EXAMPLE 4.2 The Effects of a Gasoline Tax

In part to conserve energy and in part to raise revenues, the U.S. government has often considered increasing the federal gasoline tax. In 1993, for example, a modest 7 1/2-cent increase was enacted as part of a larger budget-reform package. This increase was much less than the increase that would have been necessary to put U.S. gasoline prices on a par with those in Europe. Because an important goal of higher gasoline taxes is to discourage gasoline consumption, the government has also considered ways of passing the resulting income back to consumers. One popular suggestion is a rebate program in which tax revenues would be returned to households on an equal per capita basis. What would be the effect of such a program?

Let's begin by focusing on the effect of the program over a period of five years. The relevant price elasticity of demand is about -0.5 .¹ Suppose that a low-income consumer uses about 1200 gallons of gasoline per year, that gasoline costs \$1 per gallon, and that our consumer's annual income is \$9000.

Figure 4.9 shows the effect of the gasoline tax. (The graph has intentionally been drawn not to scale so that the effects we are discussing can be seen more clearly.) The original budget line is AB , and the consumer maximizes utility (on indifference curve U_2) by consuming the market basket at C , buying 1200 gallons of gasoline and spending \$7800 on other goods. If the tax is 50 cents per gallon, price will increase by 50 percent, shifting the new budget line to AD .² (Recall that when price changes and income stays fixed, the budget line rotates around a pivotal point on the unchanged axis.) With a price elasticity of -0.5 , consumption will decline 25 percent, from 1200 to 900 gallons, as shown by the utility-maximizing point E on indifference curve U_1 (for every 1-percent increase in the price of gasoline, quantity demanded drops by 1/2 percent).

The rebate program, however, partially counters this effect. Suppose that because the tax revenue per person is about \$450 (900 gallons times 50 cents per gallon), each consumer receives a \$450 rebate. How does this increased income affect gasoline consumption? The effect can be shown graphically by shifting the budget line upward by \$450, to line FJ , which is parallel to AD . How much gasoline does our consumer buy now? In Chapter 2, we saw that the income elasticity of demand for gasoline is approximately 0.3. Because \$450 represents a 5-percent increase in income ($\$450/\$9000 = 0.05$), we would expect the rebate to increase consumption by 1.5 percent (0.3 times 5 percent) of 900 gallons, or 13.5 gallons. The new utility-maximizing consumption choice at H reflects this expectation. (We omitted the indifference curve that is tangent at H to simplify the diagram.) Despite the rebate program, the tax would reduce gasoline consumption by 286.5 gallons, from 1200 to 913.5. Because the income elasticity of demand for gasoline is relatively low, the income effect of the rebate program is dominated by the substitution effect, and the program with a rebate does indeed reduce consumption.

In order to put a real tax-rebate program into effect, a variety of practical problems would need to be resolved. First, incoming tax receipts and rebate

¹ We saw in Chapter 2 that the price elasticity of demand for gasoline varied substantially from the short run to the long run, ranging from -0.11 in the short run to -1.17 in the long run.

² To simplify the example, we have assumed that the entire tax is paid by consumers in the form of a higher price. A broader analysis of tax shifting is presented in Chapter 9.

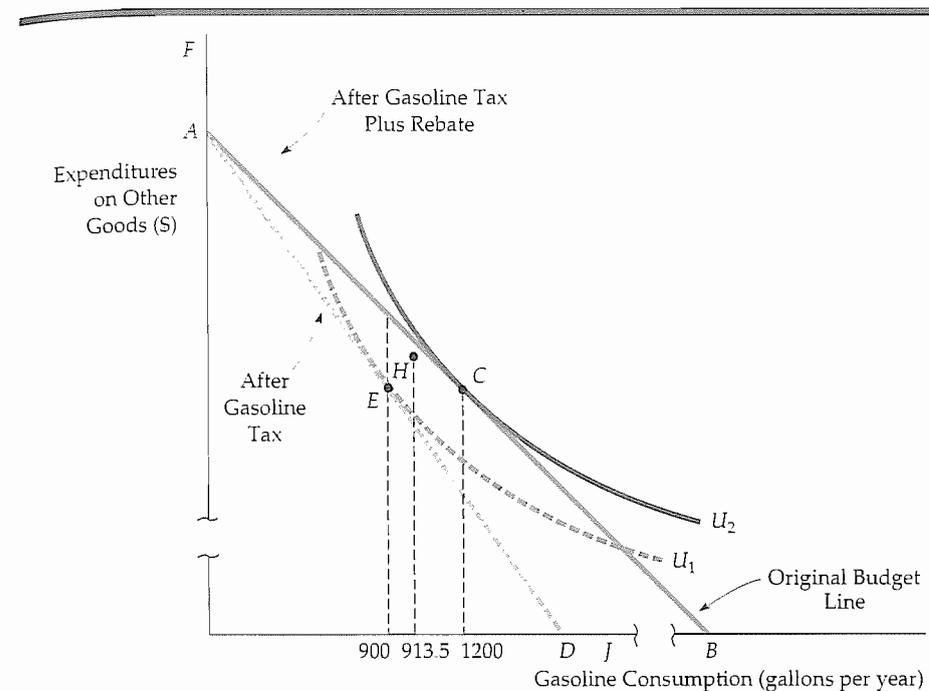


FIGURE 4.9 Effect of a Gasoline Tax with a Rebate

A gasoline tax is imposed when the consumer is initially buying 1200 gallons of gasoline at point C . After the tax takes effect, the budget line shifts from AB to AD and the consumer maximizes his preferences by choosing E , with a gasoline consumption of 900 gallons. However, when the proceeds of the tax are rebated to the consumer, his consumption increases somewhat, to 913.5 gallons at H . Despite the rebate program, the consumer's gasoline consumption has fallen, as has his level of satisfaction.

expenditures would vary from year to year, making it difficult to plan the budgeting process. For example, the tax rebate of \$450 in the first year of the program is an increase in income. During the second year, it would lead to some increase in gasoline consumption among the low-income consumers that we are studying. With increased consumption, however, the tax paid and the rebate received by this individual will increase in the second year. As a result, it may be difficult to predict the size of the program budget.

Figure 4.9 reveals that the gasoline tax program makes this particular low-income consumer slightly worse off because H lies just below indifference curve U_2 . Of course, some low-income consumers might actually benefit from the program (if, for example, they consume less gasoline on average than the group of consumers whose consumption determines the selected rebate). Nevertheless, the substitution effect caused by the tax will make consumers, on average, worse off.

Why, then, introduce such a program? Those who support gasoline taxes argue that they promote national security (by reducing dependence on foreign oil) and encourage conservation, thus helping to slow global warming by reducing the buildup of carbon dioxide in the atmosphere. We will further examine the impact of a gasoline tax in Chapter 9.

4.3 Market Demand

market demand curve Curve relating the quantity of a good that all consumers in a market will buy to its price.

So far, we have discussed the demand curve for an individual consumer. Now we turn to the **market demand curve**. Recall from Chapter 2 that the market demand curve shows how much of a good consumers overall are willing to buy as its price changes. In this section, we show how market demand curves can be derived as the sum of the individual demand curves of all consumers in a particular market.

From Individual to Market Demand

To keep things simple, let's assume that only three consumers (*A*, *B*, and *C*) are in the market for coffee. Table 4.2 tabulates several points on each consumer's demand curve. The market demand, column (5), is found by adding columns (2), (3), and (4) to determine the total quantity demanded at every price. When the price is \$3, for example, the total quantity demanded is 2 + 6 + 10, or 18.

Figure 4.10 shows these same three consumers' demand curves for coffee (labeled D_A , D_B , and D_C). In the graph, the market demand curve is the *horizontal summation* of the demands of each consumer. We sum horizontally to find the total amount that the three consumers will demand at any given price. For example, when the price is \$4, the quantity demanded by the market (11 units) is the sum of the quantity demanded by *A* (no units), by *B* (4 units), and by *C* (7 units). Because all the individual demand curves slope downward, the market demand curve will also slope downward. However, the market demand curve need not be a straight line, even though each of the individual demand curves is. In Figure 4.10, for example, the market demand curve is *kinked* because one consumer makes no purchases at prices that the other consumers find inviting (those above \$4).

Two points should be noted as a result of this analysis:

1. *The market demand curve will shift to the right as more consumers enter the market.*
2. *Factors that influence the demands of many consumers will also affect market demand.* Suppose, for example, that most consumers in a particular market earn more income and, as a result, increase their demands for coffee. Because each consumer's demand curve shifts to the right, so will the market demand curve.

TABLE 4.2 Determining the Market Demand Curve

(1) PRICE (\$)	(2) INDIVIDUAL A (UNITS)	(3) INDIVIDUAL B (UNITS)	(4) INDIVIDUAL C (UNITS)	(5) MARKET (UNITS)
1	6	10	16	32
2	4	8	13	25
3	2	6	10	18
4	0	4	7	11
5	0	2	4	6

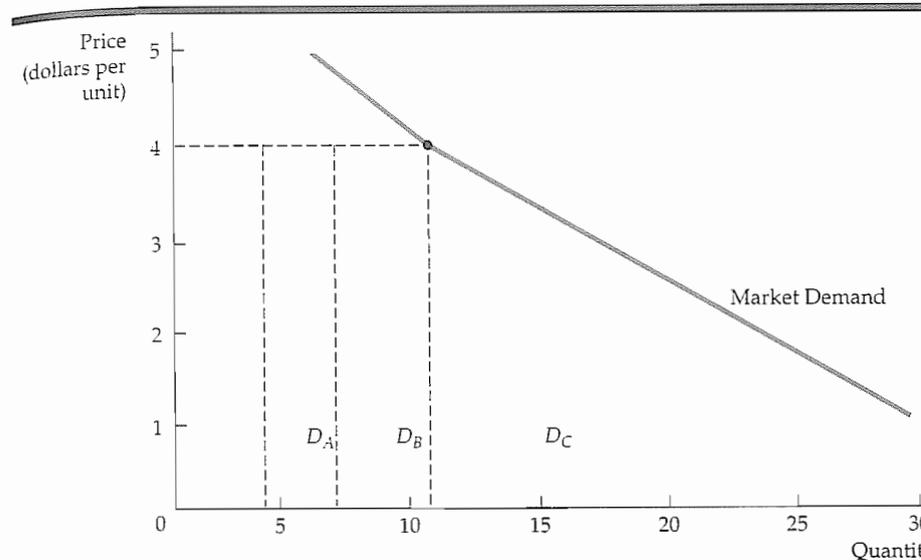


FIGURE 4.10 Summing to Obtain a Market Demand Curve

The market demand curve is obtained by summing the consumers' demand curves D_A , D_B , and D_C . At each price, the quantity of coffee demanded by the market is the sum of the quantity demanded by each consumer. At a price of \$4, for example, the quantity demanded by the market (11 units) is the sum of the quantity demanded by *A* (no units), *B* (4 units), and *C* (7 units).

The aggregation of individual demands into market demands is not just a theoretical exercise. It becomes important in practice when market demands are built up from the demands of different demographic groups or from consumers located in different areas. For example, we might obtain information about the demand for home computers by adding independently obtained information about the demands of the following groups:

- Households with children
- Households without children
- Single individuals

Or we might determine U.S. wheat demand by aggregating domestic demand (i.e., by U.S. consumers) and export demand (i.e., by foreign consumers), as we will see in Example 4.3.

Elasticity of Demand

Recall from Section 2.3 that the price elasticity of demand measures the percentage change in the quantity demanded resulting from a 1-percent change in price. Denoting the quantity of a good by Q and its price by P , the *price elasticity of demand* is

$$E_P = \frac{\Delta Q/Q}{\Delta P/P} = \left(\frac{P}{Q}\right)\left(\frac{\Delta Q}{\Delta P}\right) \quad (4.1)$$

(Here, because Δ means "a change in" $\Delta Q/Q$ is the percentage change in Q .)

In §2.3, we discuss how the price elasticity of demand describes the responsiveness of consumer demands to changes in price.

Inelastic Demand When demand is inelastic (i.e., E_p is less than 1 in magnitude), the quantity demanded is relatively unresponsive to changes in price. As a result, total expenditure on the product increases when the price increases. Suppose, for example, that a family currently uses 1000 gallons of gasoline a year when the price is \$1 per gallon; suppose also that our family's price elasticity of demand for gasoline is -0.5 . If the price of gasoline increases to \$1.10 (a 10-percent increase), the consumption of gasoline falls to 950 gallons (a 5-percent decrease). Total expenditure on gasoline, however, will increase from \$1000 (1000 gallons \times \$1 per gallon) to \$1045 (950 gallons \times \$1.10 per gallon).

Elastic Demand In contrast, when demand is elastic (E_p is greater than 1 in magnitude), total expenditure on the product decreases as the price goes up. Suppose that a family buys 100 pounds of chicken per year, at a price of \$2 per pound; the price elasticity of demand for chicken is -1.5 . If the price of chicken increases to \$2.20 (a 10-percent increase), our family's consumption of chicken falls to 85 pounds a year (a 15-percent decrease). Total expenditure on chicken will also fall, from \$200 (100 pounds \times \$2 per pound) to \$187 (85 pounds \times \$2.20 per pound).

Isoelastic Demand When the price elasticity of demand is constant all along the demand curve, we say that the curve is **isoelastic**. Figure 4.11 shows an isoelastic demand curve. Note how this demand curve is bowed inward. In contrast, recall from Section 2.3 what happens to the price elasticity of demand as we move along a *linear demand curve*. Although the slope of the linear curve is constant, the price elasticity of demand is not. It is zero when the price is zero, and it increases in magnitude until it becomes infinite when the price is sufficiently high for the quantity demanded to become zero.

isoelastic demand curve
Demand curve with a constant price elasticity.

In §2.3, we show that when the demand curve is linear, demand becomes more elastic as the price of the product increases.

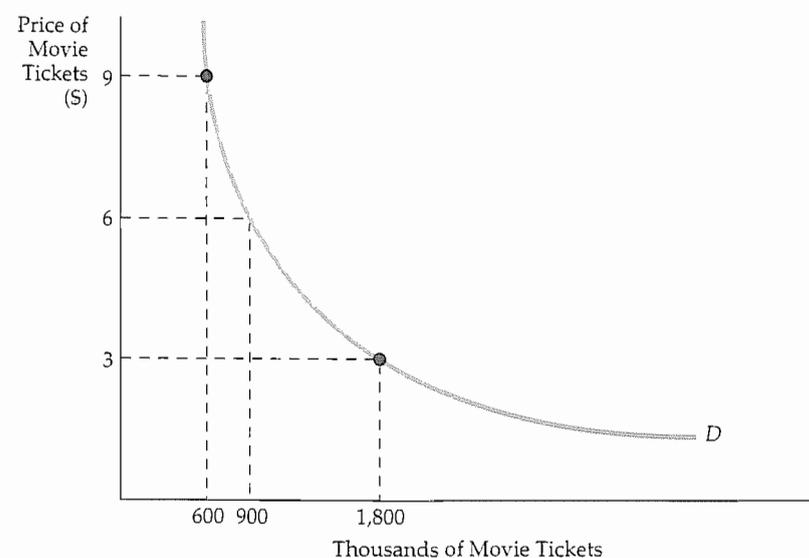


FIGURE 4.11 Unit-Elastic Demand Curve

When the price elasticity of demand is -1.0 at every price, the total expenditure is constant along the demand curve D .

TABLE 4.3 Price Elasticity and Consumer Expenditures

DEMAND	IF PRICE INCREASES, EXPENDITURES	IF PRICE DECREASES, EXPENDITURES
Inelastic	Increase	Decrease
Unit elastic	Are unchanged	Are unchanged
Elastic	Decrease	Increase

A special case of this isoelastic curve is the *unit-elastic demand curve*: a demand curve with price elasticity always equal to -1 , as is the case for the curve in Figure 4.11. In this case, total expenditure remains the same after a price change. A price increase, for instance, leads to a decrease in the quantity demanded that leaves the total expenditure on the good unchanged. Suppose, for example, that the total expenditure on first-run movies in Berkeley, California, is \$5.4 million per year, regardless of the price of a movie ticket. For all points along the demand curve, the price times the quantity will be \$5.4 million. If the price is \$6, the quantity will be 900,000 tickets; if the price increases to \$9, the quantity will drop to 600,000 tickets, as shown in Figure 4.11.

Table 4.3 summarizes the relationship between elasticity and expenditure. It is useful to review this table from the perspective of the seller of the good rather than the buyer. (What the sellers perceive as total revenue, the consumers view as total expenditures.) When demand is inelastic, a price increase leads only to a small decrease in quantity demanded; thus, the seller's total revenue increases. But when demand is elastic, a price increase leads to a large decline in quantity demanded and total revenue falls.

When calculating demand elasticities, we must be careful about the price change or quantity change in question. For a large price change (say, 20 percent), the value of the elasticity will depend on the precise point at which we measure the price and quantity along the demand curve. For this reason, it is useful to distinguish between a point elasticity of demand and an arc elasticity of demand.

Point Elasticity of Demand The **point elasticity of demand** is defined as the price elasticity at a particular point on the demand curve. Note that this is the concept of elasticity that we used throughout Chapter 2. It is calculated by substituting for $\Delta P/\Delta Q$ in the elasticity formula the magnitude of the slope of the demand curve at that point. ($\Delta P/\Delta Q$ is the slope for small ΔP because price is measured on the vertical axis and quantity demanded on the horizontal axis.) As a result, equation (4.1) becomes

point elasticity of demand
Price elasticity at a particular point on the demand curve.

$$\text{Point elasticity: } E_p = (P/Q)(1/\text{slope}) \quad (4.2)$$

There are times when we want to calculate a price elasticity over some portion of the demand curve rather than at a single point. Suppose, for example, that we are contemplating an increase in the price of a product from \$8 to \$10 and expect the quantity demanded to fall from 6 units to 4. How should we calculate the price elasticity of demand? Is the price increase 25 percent (a \$2 increase divided by the original price of \$8), or is it 20 percent (a \$2 increase divided by the new price of \$10)? Is the percentage decrease in quantity demanded 33 1/3 percent (2/6) or 50 percent (2/4)?

There is no correct answer to such questions. We could calculate the price elasticity using the original price and quantity. If so, we would find that $E_p = (-33\frac{1}{3}\text{ percent}/25\text{ percent}) = -1.33$. Or we could use the new price and quantity, in which case we would find that $E_p = (-50\text{ percent}/20\text{ percent}) = -2.5$. The difference between these two calculated elasticities is large, and neither seems preferable to the other.

arc elasticity of demand Price elasticity calculated over a range of prices.

Arc Elasticity of Demand We can resolve this problem by using the arc elasticity of demand: the elasticity calculated over a range of prices. Rather than choose either the initial or the final price, we use an average of the two, \bar{P} ; for the quantity demanded, we use \bar{Q} . Thus the arc elasticity of demand is given by

$$\text{Arc elasticity: } E_p = (\Delta Q/\Delta P)(\bar{P}/\bar{Q}) \quad (4.3)$$

In our example, the average price is \$9 and the average quantity 5 units. Thus the arc elasticity is

$$E_p = (-2/\$2)(\$9/5) = -1.8$$

The arc elasticity will always lie somewhere (but not necessarily halfway) between the point elasticities calculated at the lower and the higher prices.

Although the arc elasticity of demand is sometimes useful, economists generally use the word "elasticity" to refer to a *point* elasticity. Throughout the rest of this book, we will do the same, unless noted otherwise.

EXAMPLE 4.3 The Aggregate Demand for Wheat

In Chapter 2 (Example 2.4), we explained that the demand for U.S. wheat has two components: domestic demand (by U.S. consumers) and export demand (by foreign consumers). Let's see how the total demand for wheat during 1998 can be obtained by aggregating the domestic and foreign demands.

Domestic demand for wheat is given by the equation

$$Q_{DD} = 1700 - 107P$$

where Q_{DD} is the number of bushels (in millions) demanded domestically, and P is the price in dollars per bushel. Export demand is given by

$$Q_{DE} = 1544 - 176P$$

where Q_{DE} is the number of bushels (in millions) demanded from abroad. As shown in Figure 4.12, domestic demand, given by AB , is relatively price inelastic. (Statistical studies have shown that price elasticity of domestic demand is about -0.2 .) However, export demand, given by CD , is more price elastic, with

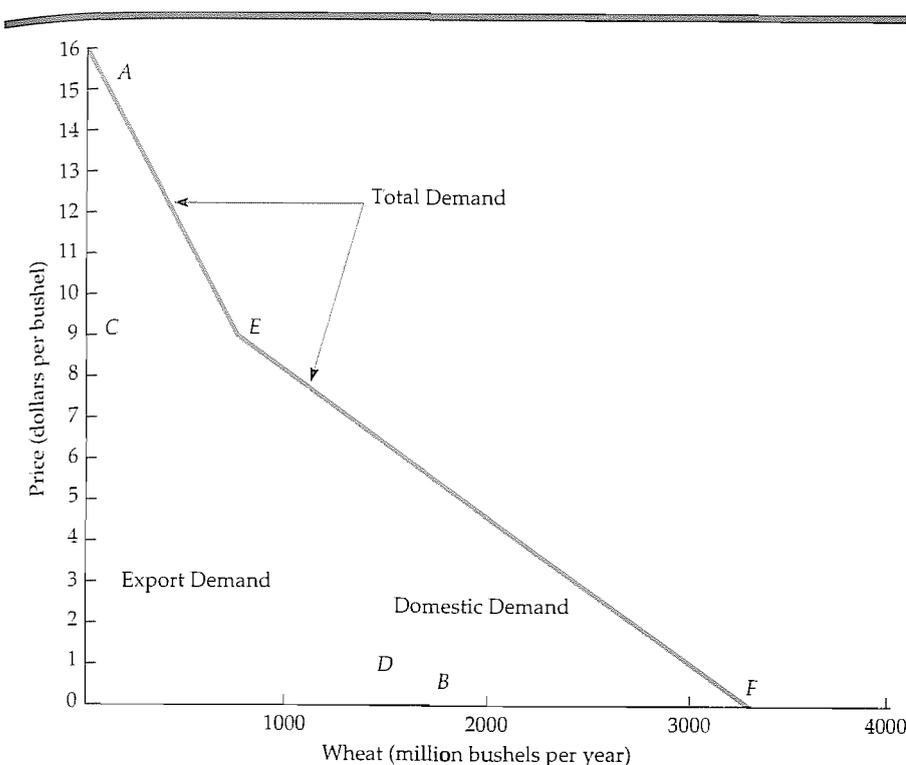


FIGURE 4.12 The Aggregate Demand for Wheat

The total world demand for wheat is the horizontal sum of the domestic demand AB and the export demand CD . Even though each individual demand curve is linear, the market demand curve is kinked, reflecting the fact that there is no export demand when the price of wheat is greater than about \$9 per bushel.

an elasticity of -0.4 . Export demand is more elastic than domestic demand because poorer countries that import U.S. wheat turn to other grains and foodstuffs if wheat prices rise.³

To obtain the world demand for wheat, we set the left side of each demand equation equal to the quantity of wheat (the variable on the horizontal axis). We then add the right side of the equations, obtaining

$$Q_{DD} + Q_{DE} = (1700 - 107P) + (1544 - 176P) = 3244 - 283P$$

(Note that this is the same total demand equation for 1998 given in Example 2.4.) This generates the line segment EF in Figure 4.12.

³ For a survey of statistical studies of demand and supply elasticities and an analysis of the U.S. wheat market, see Larry Salathe and Sudchada Langley, "An Empirical Analysis of Alternative Export Subsidy Programs for U.S. Wheat," *Agricultural Economics Research* 38, No. 1 (Winter 1986).

At all prices above point C, however, there is no export demand, so that world demand and domestic demand are identical. As a result, for all prices above C, world demand is given by line segment AE. (If we were to add Q_{DE} for prices above C, we would be incorrectly adding a negative export demand to a positive domestic demand.) As the figure shows, the resulting total demand for wheat, given by AEF, is kinked. The kink occurs at point E, the price level above which there is no export demand.

EXAMPLE 4.4 The Demand for Housing

Several years ago, the U.S. Department of Housing and Urban Development began an experimental program of housing subsidies, a program aimed at easing the housing burdens of the poor. The subsidies typically involved supplements based solely on income but could alternatively have been designed as a percentage of housing expenditures. In order to determine the effects of such a program on various demographic groups, we need information about the price and income elasticities of demand for housing.

A family's demand for housing depends on the age and status of the household making the purchasing decision. One approach to housing demand is to relate the number of rooms per house for each household (the quantity demanded) both to an estimate of the price of an additional room in a house and to the household's family income.⁴ (Prices of rooms vary because of differences in construction costs.) Table 4.4 lists some of the price and income elasticities for different demographic groups.

In general, the elasticities show that the size of houses that consumers demand (as measured by the number of rooms) is relatively insensitive to differences in either income or price. However, differences do appear among subgroups of the population. For example, families with young household heads have a price elasticity of -0.22 , which is substantially greater than those with

TABLE 4.4 Price and Income Elasticities of the Demand for Rooms

GROUP	PRICE ELASTICITY	INCOME ELASTICITY
Single individuals	-0.14	0.19
Married, head of household age less than 30, 1 child	-0.22	0.07
Married, head age 30-39, 2 or more children	0	0.11
Married, head age 50 or older, 1 child	-0.08	0.18

⁴ See Mahlon Strazheim, *An Econometric Analysis of the Urban Housing Market* (New York: National Bureau of Economic Research, 1975), ch. 4.

older household heads. Presumably, families buying houses are more price sensitive when parents and their children are younger and parents may be planning for more children. Among married households, the income elasticity of demand for rooms also increases with age—a fact which tells us that older households buy larger houses than younger households.

Price and income elasticities of demand for housing also depend on where people live.⁵ Demand in central cities is much more price elastic than in suburbs. Income elasticities, however, increase as one moves farther from the central city. Thus poorer (on average) central-city residents (who live where the price of land is relatively high) are more price sensitive in their housing choices than their wealthier suburban counterparts.

4.4 Consumer Surplus

Consumers buy goods because the purchase makes them better off. **Consumer surplus** measures *how much* better off individuals are, in the aggregate, because they can buy goods in the market. Because different consumers place different values on the consumption of particular goods, the maximum amount they are willing to pay for those goods also differs. *Consumer surplus is the difference between the maximum amount that a consumer is willing to pay for a good and the amount that the consumer actually pays.* Suppose, for example, that a student would have been willing to pay \$13 for a rock concert ticket even though she had to pay only \$12. The \$1 difference is her consumer surplus.⁶ When we add the consumer surpluses of all consumers who buy a good, we obtain a measure of the aggregate consumer surplus.

(individual) consumer surplus Difference between what a consumer is willing to pay for a good and the amount actually paid.

Consumer Surplus and Demand

Consumer surplus can be calculated easily if we know the demand curve. To see the relationship between demand and consumer surplus, let's examine the individual demand curve for concert tickets shown in Figure 4.13. (Although the following discussion applies to an individual demand curve, a similar argument also applies to a market demand curve.) Drawing the demand curve as a staircase rather than a straight line shows us how to measure the value that our consumer obtains from buying different numbers of tickets.

When deciding how many tickets to buy, our student might reason as follows: The first ticket costs \$14 but is worth \$20. This \$20 valuation is obtained by using the demand curve to find the maximum amount that she will pay for each *additional* ticket (\$20 being the maximum that she will pay for the *first* ticket). The first ticket is worth purchasing because it generates \$6 of surplus value above and beyond its cost. The second ticket is also worth buying because it generates a surplus of \$5 (\$19 - \$14). The third ticket generates a surplus of \$4. The

⁵ See Allen C. Goodman and Masahiro Kawai, "Functional Form, Sample Selection, and Housing Demand," *Journal of Urban Economics* 20 (September 1986): 155-67.

⁶ Measuring consumer surplus in dollars involves an implicit assumption about the shape of consumers' indifference curves: namely, that the marginal utility associated with increases in a consumer's income remains constant within the range of income in question. In many cases, this is a reasonable assumption. It may be suspect, however, when large changes in income are involved.

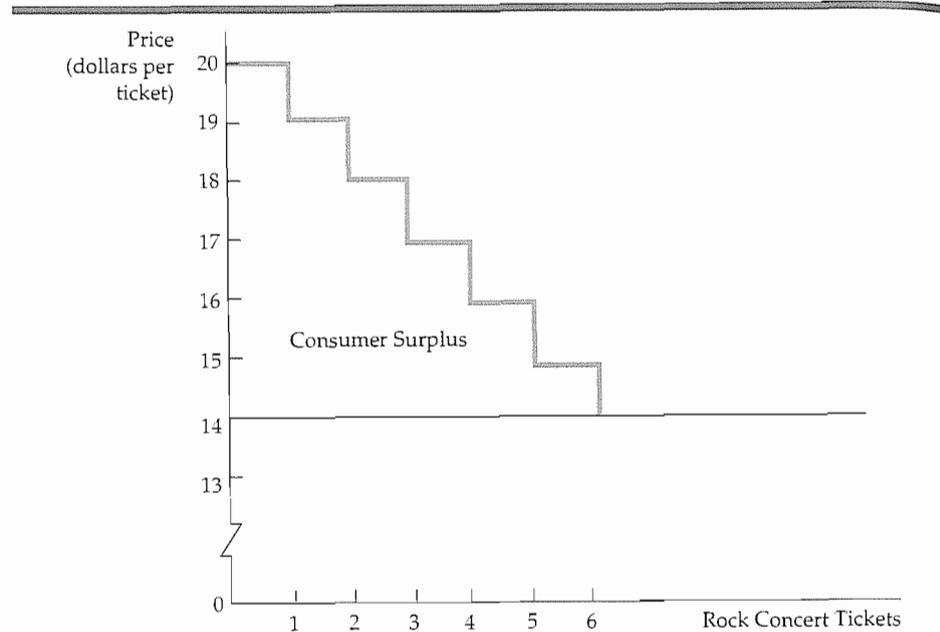


FIGURE 4.13 Consumer Surplus

Consumer surplus is the total benefit from the consumption of a product, net of the total cost of purchasing it. Here, the consumer surplus associated with six concert tickets (purchased at \$14 per ticket) is given by the yellow-shaded area.

fourth, however, generates a surplus of only \$3, the fifth a surplus of \$2, and the sixth a surplus of just \$1. Our student is indifferent about purchasing the seventh ticket (which generates zero surplus) and prefers not to buy any more than that because the value of each additional ticket is less than its cost. In Figure 4.13, consumer surplus is found by *adding the excess values or surpluses for all units purchased*. In this case, then, consumer surplus equals

$$\$6 + \$5 + \$4 + \$3 + \$2 + \$1 = \$21$$

To calculate the aggregate consumer surplus in a market, we simply find the area below the *market* demand curve and above the price line. For our rock concert example, this principle is illustrated in Figure 4.14. Now, because the number of tickets sold is measured in thousands and individual demand curves differ, the market demand curve appears as a straight line. Note that the actual expenditure on tickets is $6500 \times \$14 = \$91,000$. Consumer surplus, shown as the shaded triangle, is

$$1/2 \times (\$20 - \$14) \times 6500 = \$19,500$$

This figure is the total benefit to consumers, less what they paid for the tickets.

Of course market demand curves are not always straight lines. Nonetheless, we can always measure consumer surplus by finding the area below the demand curve and above the price line.

Applying Consumer Surplus Consumer surplus has important applications in economics. When added over many individuals, it measures the aggregate

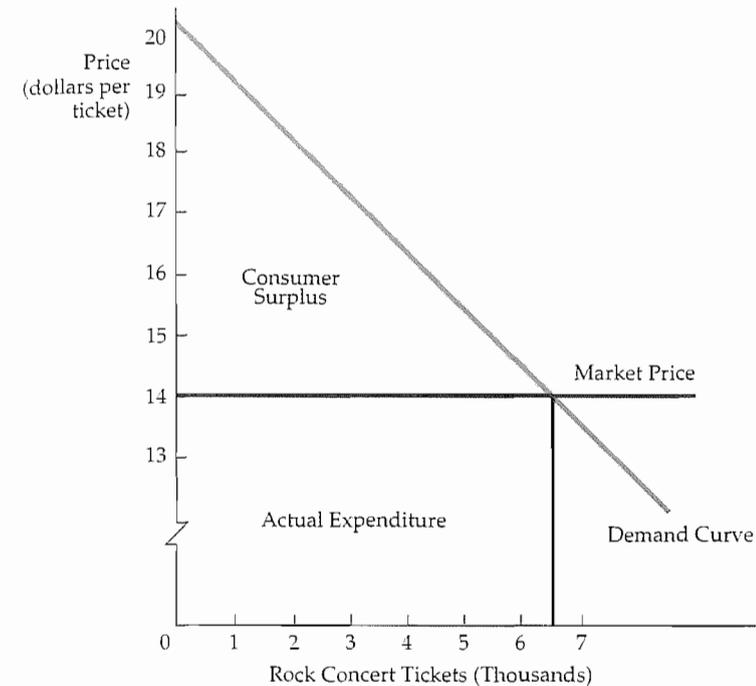


FIGURE 4.14 Consumer Surplus Generalized

For the market as a whole, consumer surplus is measured by the area under the demand curve and above the line representing the purchase price of the good. Here, the consumer surplus is given by the shaded triangle and is equal to $1/2 \times (\$20 - \$14) \times 6500 = \$19,500$.

benefit that consumers obtain from buying goods in a market. When we combine consumer surplus with the aggregate profits that producers obtain, we can evaluate both the costs and benefits not only of alternative market structures, but of public policies that alter the behavior of consumers and firms in those markets.

EXAMPLE 4.5 The Value of Clean Air

Air is free in the sense that we don't pay to breathe it. But the absence of a market for air may help explain why the air quality in some cities has been deteriorating for decades. To encourage cleaner air, Congress passed the Clean Air Act in 1963 and has since amended it a number of times. In 1970, for example, automobile emissions controls were tightened. Were these controls worth it? Were the benefits of cleaning up the air sufficient to outweigh the costs imposed directly on car producers and indirectly on car buyers?

To answer this question, Congress asked the National Academy of Sciences to evaluate emissions controls in a cost-benefit study. Using empirically determined estimates of the demand for clean air, the benefits portion of the study determined how much people value clean air. Although there is no actual market for clean air, people do pay more for houses where the air is clean than for

comparable houses in areas with dirtier air. This information was used to estimate the demand for clean air.⁷ Detailed data on house prices in neighborhoods of Boston and Los Angeles were compared with the levels of various air pollutants. The effects of other variables that might affect house value were taken into account statistically. The study determined a demand curve for clean air that looked approximately like the one shown in Figure 4.15.

The horizontal axis measures the amount of air *pollution reduction*; the vertical axis measures the increased value of a home associated with those reductions. Consider, for example, the demand for cleaner air of a homeowner in a city in which the air is rather dirty, as exemplified by a level of nitrogen oxides (NOX) of 10 parts per 100 million (pphm). If the family were required to pay \$1000 for each 1 pphm reduction in air pollution, it would choose *A* on the demand curve in order to obtain a pollution reduction of 5 pphm.

How much is a 50-percent, or 5-pphm, reduction in pollution worth to the same family? We can measure this value by calculating the consumer surplus associated with reducing air pollution. Because the price for this reduction is \$1000 per unit, the family would pay \$5000. However, the family values all but the last unit of reduction by more than \$1000. As a result, the shaded triangle in Figure 4.15 gives the value of the cleanup (above and beyond the payment). Because the demand curve is a straight line, the surplus can be calculated from the area of the triangle whose height is \$1000 (\$2000 – \$1000) and whose base is 5 pphm. Therefore, the value to the household of the pollution reduction is \$2500.

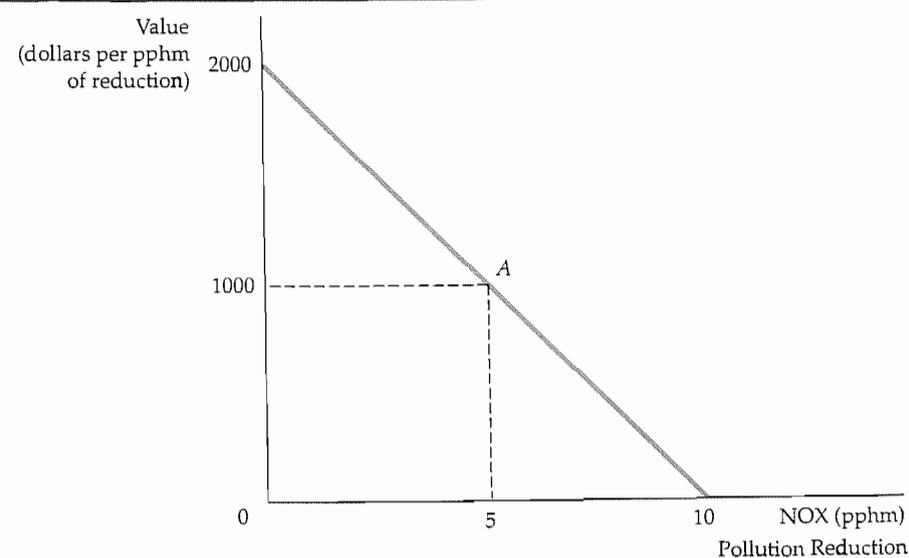


FIGURE 4.15 Valuing Cleaner Air

The shaded triangle gives the consumer surplus generated when air pollution is reduced by 5 parts per 100 million of nitrogen oxide at a cost of \$1000 per part reduced. The surplus is created because most consumers are willing to pay more than \$1000 for each unit reduction of nitrogen oxide.

⁷ The results are summarized in Daniel L. Rubinfeld, "Market Approaches to the Measurement of the Benefits of Air Pollution Abatement," in Ann Friedlaender, ed., *The Benefits and Costs of Cleaning the Air* (Cambridge: MIT Press, 1976), 240–73.

A complete cost-benefit analysis would use a measure of the total benefit of the cleanup—the benefit per household times the number of households. This figure could be compared with the total cost of the cleanup to determine whether such a project was worthwhile. We will discuss clean air further in Chapter 18, when we describe the tradeable emissions permits that were introduced by the Clean Air Act Amendments of 1990.

4.5 Network Externalities

So far, we have assumed that people's demands for a good are independent of one another. In other words, Tom's demand for coffee depends on Tom's tastes and income, the price of coffee, and perhaps the price of tea. But it does not depend on Dick's or Harry's demands for coffee. This assumption has enabled us to obtain the market demand curve simply by summing individuals' demands.

For some goods, however, one person's demand also depends on the demands of *other* people. In particular, a person's demand may be affected by the number of other people who have purchased the good. If this is the case, there exists a **network externality**. Network externalities can be positive or negative. A *positive* network externality exists if the quantity of a good demanded by a typical consumer increases in response to the growth in purchases of other consumers. If the quantity demanded decreases, there is a *negative* network externality.

network externality When each individual's demand depends on the purchases of other individuals.

The Bandwagon Effect

One example of a positive network externality is the **bandwagon effect**—the desire to be in style, to possess a good because almost everyone else has it, or to indulge in a fad.⁸ The bandwagon effect often arises with children's toys (Beanie Babies or Sega video games, for example). In fact, exploiting this effect is a major objective in marketing and advertising toys. Often it is also the key to success in selling clothing.

bandwagon effect Positive network externality in which a consumer wishes to possess a good in part because others do.

The bandwagon effect is illustrated in Figure 4.16, in which the horizontal axis measures the sales of some fashionable good in thousands per month. Suppose consumers think that only 20,000 people have bought a certain good. Because this is a small number relative to the total population, consumers will have little motivation to buy the good in order to be in style. Some consumers may still buy it (depending on price), but only for its intrinsic value. In this case, demand is given by the curve D_{20} .

Suppose instead that consumers think that 40,000 people have bought the good. Now they find the good more attractive and want to buy more. The demand curve is D_{40} , which is to the right of D_{20} . Similarly, if consumers think that 60,000 people have bought the good, the demand curve will be D_{60} , and so on. The more people consumers believe to have bought the good, the farther to the right the demand curve shifts.

⁸ The bandwagon effect and the snob effect were introduced by Harvey Liebenstein, "Bandwagon, Snob, and Veblen Effects in the Theory of Consumers' Demand," *Quarterly Journal of Economics* 62 (February 1948): 165–201.

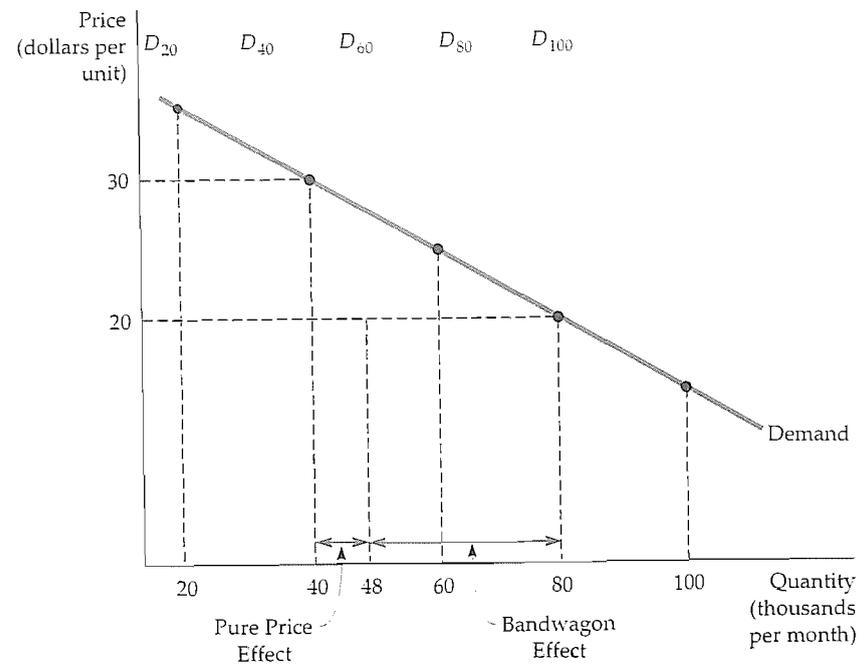


FIGURE 4.16 Positive Network Externality: Bandwagon Effect

A bandwagon effect is a positive network externality in which the quantity of a good that an individual demands grows in response to the growth of purchases by other individuals. Here, as the price of the product falls from \$30 to \$20, the bandwagon effect causes the demand for a good to shift to the right, from D_{40} to D_{80} .

Ultimately, consumers will get a good sense of how many people have in fact purchased a good. This number will depend, of course, on its price. In Figure 4.16, for example, we see that if the price were \$30, 40,000 people would buy the good. Thus the relevant demand curve would be D_{40} . If the price were \$20, 80,000 people would buy the good and the relevant demand curve would be D_{80} . The market demand curve is therefore found by joining the points on the curves D_{20} , D_{40} , D_{60} , D_{80} , and D_{100} that correspond to the quantities 20,000, 40,000, 60,000, 80,000 and 100,000.

Compared with the curves D_{20} , etc., the market demand curve is relatively elastic. To see why the bandwagon effect leads to a more elastic demand curve, consider the effect of a drop in price from \$30 to \$20, with a demand curve of D_{40} . If there were no bandwagon effect, quantity demanded would increase from 40,000 to only 48,000. But as more people buy the good and it becomes stylish to own it, the bandwagon effect increases quantity demanded further, to 80,000. Thus the bandwagon effect increases the response of demand to price changes—i.e., it makes demand more elastic. As we'll see later, this result has important implications for producers' pricing strategies.

Although the bandwagon effect is associated with fads and stylishness, positive network externalities can arise for other reasons. The greater the number of people who own a particular good, the greater the intrinsic value of that good to each owner. For example, if I am the only person to own a compact disc player, it will not be economical for companies to manufacture compact discs; without the discs, the CD player will obviously be of little value to me. But the greater the

number of people who own players, the more discs will be manufactured and the greater will be the value of the player to me. The same is true for personal computers: The more people who own them, the more software will be written, and thus the more useful the computer will be to me.

The Snob Effect

Network externalities are sometimes negative. Consider the **snob effect**, which refers to the desire to own exclusive or unique goods. The quantity demanded of a "snob good" is higher the *fewer* the people who own it. Rare works of art, specially designed sports cars, and made-to-order clothing are snob goods. The value one gets from a painting or a sports car is partly the prestige, status, and exclusivity resulting from the fact that few other people own one like it.

Figure 4.17 illustrates the snob effect. D_2 is the demand curve that would apply if consumers believed that only 2,000 people owned the good. If they believe that 4,000 people own the good, it is less exclusive, and so its snob value is reduced. Quantity demanded will therefore be lower; the curve D_4 applies. Similarly, if consumers believe that 6,000 people own the good, demand is even smaller and D_6 applies. Eventually, consumers learn how widely owned a good

snob effect Negative network externality in which a consumer wishes to own an exclusive or unique good.

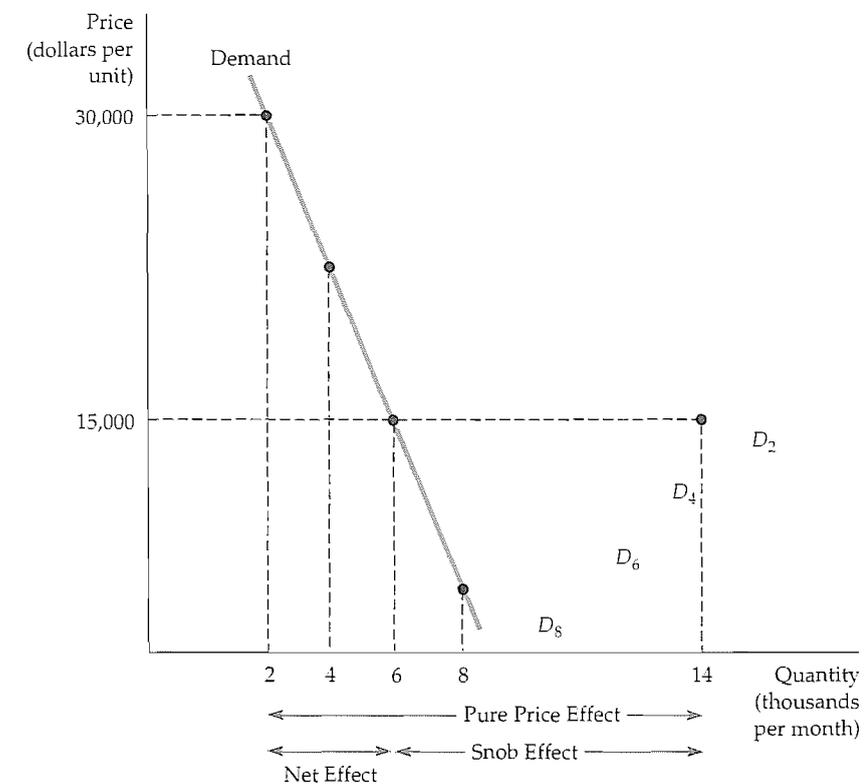


FIGURE 4.17 Negative Network Externality: Snob Effect

A snob effect is a negative network externality in which the quantity of a good that an individual demands falls in response to the growth of purchases by other individuals. Here, as the price falls from \$30,000 to \$15,000 and more people buy the good, the snob effect causes the demand for a good to shift to the left, from D_2 to D_6 .

actually is. Thus the market demand curve is found by joining the points on the curves D_2 , D_4 , D_6 , etc., that actually correspond to the quantities 2000, 4000, 6000, etc.

The snob effect makes market demand less elastic. To see why, suppose the price was initially \$30,000, with 2000 people purchasing the good. What happens when the price is lowered to \$15,000? If there were no snob effect, the quantity purchased would increase to 14,000 (along curve D_2). But as a snob good, its value is greatly reduced if more people own it. The snob effect dampens the increase in quantity demanded, cutting it by 8000 units; the net increase in sales is only to 6000 units. For many goods, marketing and advertising are geared to creating a snob effect (e.g., Rolex watches). The goal is less elastic demand—a result that makes it possible for firms to raise price.

Negative network externalities can arise for other reasons. Consider the effect of congestion. Because I prefer short lines and fewer skiers on the slopes, the value I obtain from a lift ticket at a ski resort is lower the more people there are who have bought tickets. Likewise for entry to an amusement park, skating rink, or beach.⁹

EXAMPLE 4.6 Network Externalities and the Demands for Computers and E-Mail

The 1950s and 1960s witnessed phenomenal growth in the demand for mainframe computers. From 1954 to 1965, for example, annual revenues from the leasing of mainframes increased at the extraordinary rate of 78 percent per year, while prices declined by 20 percent per year. Granted, prices were falling, and the quality of computers was also increasing dramatically, but the elasticity of demand would have to have been quite large to account for this kind of growth. IBM, among other computer manufacturers, wanted to know what was going on.

An econometric study by Gregory Chow helped provide some answers.¹⁰ Chow found that the demand for computers follows a “saturation curve”—a dynamic process whereby demand, though small at first, grows slowly. Soon, however, it grows rapidly, until finally nearly everyone likely to buy a product has done so, whereby the market becomes saturated. This rapid growth occurs because of a positive network externality: As more and more organizations own computers, as more and better software is written, and as more people are trained to use computers, the value of having a computer increases. Because this process causes demand to increase, still more software and better trained users are needed, and so on.

This network externality was an important part of the demand for computers. Chow found that it could account for nearly half the rapid growth of rentals between 1954 and 1965. Reductions in the inflation-adjusted price (he found a price elasticity of demand for computers of -1.44) and major increases in power and quality, which also made them much more useful and effective,

⁹ Tastes, of course, differ. Some people associate a *positive* network externality with skiing or a day on the beach; they enjoy crowds and may even find the slope or beach lonely without them.

¹⁰ See Gregory Chow, “Technological Change and the Demand for Computers,” *American Economic Review* 57, no. 5 (December 1967): 1117–30.

accounted for the other half. Other studies have shown that this process continued through the following decades.¹¹ In fact, this same kind of network externality helped to fuel a rapid growth rate in the demand for personal computers.

Today there is little debate about the importance of network externalities as an explanation for the success of Microsoft’s Windows PC operating system, which by 1999 was being used in about 90 percent of personal computers worldwide. At least as significant has been the phenomenal success of the Microsoft Office Suite of PC applications (which includes Word and Excel). In 1999, Microsoft Office had well over 90 percent of the market.

Network externalities are not limited to computers. In recent years there has been explosive growth in the use of e-mail. Clearly a strong positive network externality is at work. Because an e-mail can only be transmitted to another e-mail user, the value of using e-mail depends crucially on how many other people use it. By the mid-1990s, nearly all business offices in the United States used e-mail, and e-mail had become a standard means of communicating.

*4.6 Empirical Estimation of Demand

Later in this book, we explain how demand information is used as input into a firm’s economic decision-making process. General Motors, for example, must understand automobile demand to decide whether to offer rebates or below-market-rate loans for new cars. Knowledge about demand is also important for public policy decisions. Understanding the demand for oil, for instance, can help Congress decide whether to pass an oil import tax. You may wonder how it is that economists determine the shape of demand curves and how price and income elasticities of demand are actually calculated. In this starred section, we will briefly examine some methods for evaluating and forecasting demand. The section is starred not only because the material is more advanced, but also because it is not needed for much of the later analysis in the book. Nonetheless, this material is instructive and will help you appreciate the empirical foundation of the theory of consumer behavior. The basic statistical tools for estimating demand curves and demand elasticities are described in the appendix to this book.

Interview and Experimental Approaches to Demand Determination

One way to obtain information about demand is through *interviews* in which consumers are asked how much of a product they might be willing to buy at a given price. This approach, however, may not succeed when people lack information or interest or even want to mislead the interviewer. Therefore, market researchers have designed various indirect survey techniques. Consumers might be asked, for example, what their current consumption behavior is and how they would respond if a certain product were available at, say, a 10-percent discount.

¹¹ See Robert J. Gordon, “The Postwar Evolution of Computer Prices,” in Dale W. Jorgenson and Ralph Landau, eds., *Technology and Capital Formation* (Cambridge: MIT Press, 1989).

They might be asked how they would expect others to behave. Although indirect approaches to demand estimation can be fruitful, the difficulties of the interview approach have forced economists and marketing specialists to look to alternative methods.

In *direct marketing experiments*, actual sales offers are posed to potential customers. An airline, for example, might offer a reduced price on certain flights for six months, partly to learn how the price change affects demand for flights and partly to learn how competitors will respond.

Direct experiments are real, not hypothetical, but even so, problems remain. The wrong experiment can be costly, and even if profits and sales rise, the firm cannot be entirely sure that these increases resulted from the experimental change; other factors probably changed at the same time. Moreover, the response to experiments—which consumers often recognize as short-lived—may differ from the response to permanent changes. Finally, a firm can afford to try only a limited number of experiments.

The Statistical Approach to Demand Estimation

Firms often rely on market data based on actual studies of demand. Properly applied, the statistical approach to demand estimation can help researchers sort out the effects of variables, such as income and the prices of other products, on the quantity of a product demanded. Here we outline some of the conceptual issues involved in the statistical approach.

Table 4.5 shows the quantity of raspberries sold in a market each year. Information about the market demand for raspberries would be valuable to an organization representing growers because it would allow them to predict sales on the basis of their own estimates of price and other demand-determining variables. Let's suppose that, focusing on demand, researchers find that the quantity of raspberries produced is sensitive to weather conditions but not to the current market price (because farmers make their planting decisions based on last year's price).

The price and quantity data from Table 4.5 are graphed in Figure 4.18. If we believe that price alone determines demand, it would be plausible to describe the demand for the product by drawing a straight line (or other appropriate curve), $Q = a - bP$, which "fit" the points as shown by demand curve D . (The "least-squares" method of curve-fitting is described in the appendix to this book.)

TABLE 4.5 Demand Data			
YEAR	QUANTITY (Q)	PRICE (P)	INCOME (I)
1988	4	24	10
1989	7	20	10
1990	8	17	10
1991	13	17	17
1992	16	10	17
1993	15	15	17
1994	19	12	20
1995	20	9	20
1996	22	5	20

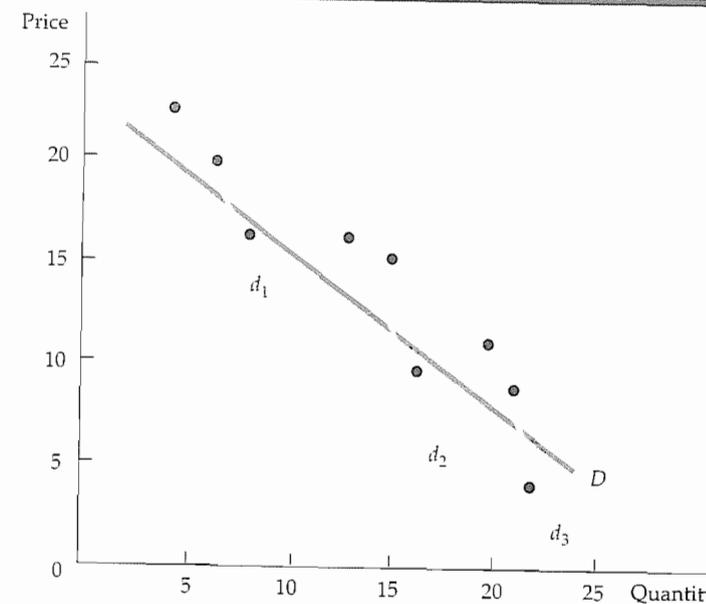


FIGURE 4.18 Estimating Demand

Price and quantity data can be used to determine the form of a demand relationship. But the same data could describe a single demand curve D or three demand curves d_1 , d_2 , and d_3 that shift over time.

Does curve D (given by the equation $Q = 28.2 - 1.00P$) really represent the demand for the product? The answer is yes—but only if no important factors other than product price affect demand. In Table 4.5, however, we have included data for one other variable: the average income of purchasers of the product. Note that income (I) has increased twice during the study, suggesting that the demand curve has shifted twice. Thus demand curves d_1 , d_2 , and d_3 in Figure 4.18 give a more likely description of demand. This demand relationship would be described algebraically as

$$Q = a - bP + cI \tag{4.4}$$

The income term in the demand equation allows the demand curve to shift in a parallel fashion as income changes. (The demand relationship, calculated using the least-squares method, is given by $Q = 8.08 - .49P + .81I$.)

The Form of the Demand Relationship

Because the demand relationships discussed above are straight lines, the effect of a change in price on quantity demanded is constant. However, the price elasticity of demand varies with the price level. For the demand equation $Q = a - bP$, for example, the price elasticity E_p is

$$E_p = (\Delta Q/\Delta P)(P/Q) = -b(P/Q) \tag{4.5}$$

Thus elasticity increases in magnitude as the price increases (and the quantity demanded falls).

There is no reason to expect elasticities of demand to be constant. Nevertheless, we often find the *isoelastic demand curve*, in which the price elasticity and the income elasticity are constant, useful to work with. When written in its *log-linear form*, the isoelastic demand curve appears as follows:

$$\log(Q) = a - b \log(P) + c \log(I) \quad (4.6)$$

where $\log(\)$ is the logarithmic function and a , b , and c are the constants in the demand equation. The appeal of the log-linear demand relationship is that the slope of the line $-b$ is the price elasticity of demand and the constant c is the income elasticity.¹² Using the data in Table 4.5, for example, we obtained the regression line

$$\log(Q) = -0.81 - 0.24 \log(P) + 1.46 \log(I)$$

This relationship tells us that the price elasticity of demand for raspberries is -0.24 (that is, demand is inelastic) and that the income elasticity is 1.46 .

We have seen that it can be useful to distinguish between goods that are complements and goods that are substitutes. Suppose that P_2 represents the price of a second good—one which is believed to be related to the product we are studying. We can then write the demand function in the following form:

$$\log(Q) = a - b \log(P) + b_2 \log(P_2) + c \log(I)$$

When b_2 , the cross-price elasticity, is positive, the two goods are substitutes; when b_2 is negative, the two goods are complements.

EXAMPLE 4.7 The Demand for Ready-to-Eat Cereal

The Post Cereals Division of Kraft General Foods acquired the Shredded Wheat cereals of Nabisco in 1995. The acquisition raised the legal and economic question of whether Post would raise the price of its best-selling brand, Grape Nuts, or the price of Nabisco's most successful brand, Shredded Wheat Spoon Size.¹³ One important issue in a lawsuit brought by the state of New York was whether the two brands were close substitutes for one another. If so, it would be more profitable for Post to increase the price of Grape Nuts after

¹² The natural logarithmic function with base e has the property that $\Delta(\log(Q)) = \Delta Q/Q$ for any change in $\log(Q)$. Similarly, $\Delta(\log(P)) = \Delta P/P$ for any change in $\log(P)$. It follows that $\Delta(\log(Q)) = \Delta Q/Q = -b[\Delta(\log(P))] = -b(\Delta P/P)$. Therefore, $(\Delta Q/Q)/(\Delta P/P) = -b$, which is the price elasticity of demand. By a similar argument, the income elasticity of demand c is given by $(\Delta Q/Q)/(\Delta I/I)$.

¹³ *State of New York v. Kraft General Foods, Inc.*, 926 F. Supp. 321, 356 (S.D.N.Y. 1995).

rather than before the acquisition. Why? Because after the acquisition the lost sales from consumers who would switch away from Grape Nuts would be recovered to the extent that they switched to Shredded Wheat.

The extent to which a price increase will cause consumers to switch is given (in part) by the price elasticity of demand for Grape Nuts. Other things being equal, the higher the demand elasticity, the greater the loss of sales associated with a price increase. The more likely, too, that the price increase will be unprofitable.

The substitutability of Grape Nuts and Shredded Wheat can be measured by the cross-price elasticity of demand for Grape Nuts with respect to the price of Shredded Wheat. The relevant elasticities were calculated using weekly data obtained from the supermarket scanning of household purchases for 10 cities over a three-year period. One of the estimated isoelastic demand equations appeared in the following log-linear form:

$$\log(Q_{GN}) = 1.998 - 2.085 \log(P_{GN}) + 0.62 \log(I) + 0.14 \log(P_{SW})$$

where Q_{GN} is the amount (in pounds) of Grape Nuts sold weekly, P_{GN} the price per pound of Grape Nuts, I real personal income, and P_{SW} the price per pound of Shredded Wheat Spoon Size.

The demand for Grape Nuts is elastic (at current prices), with a price elasticity of about -2 . The income elasticity is 0.62 : In other words, increases in income lead to increases in cereal purchases, but at less than a 1-for-1 rate. Finally, the cross-price elasticity is 0.14 . This figure is consistent with the fact that although the two cereals are substitutes (the quantity demanded of Grape Nuts increases in response to an increase in the price of Shredded Wheat), they are not very close substitutes.

SUMMARY

- Individual consumers' demand curves for a commodity can be derived from information about their tastes for all goods and services and from their budget constraints.
- Engel curves, which describe the relationship between the quantity of a good consumed and income, can be useful for discussions of how consumer expenditures vary with income.
- Two goods are substitutes if an increase in the price of one leads to an increase in the quantity demanded of the other. In contrast, two goods are complements if an increase in the price of one leads to a decrease in the quantity demanded of the other.
- The effect of a price change on the quantity demanded of a good can be broken into two parts: a substitution effect, in which satisfaction remains constant while price changes, and an income effect, in which the price remains constant while satisfaction changes. Because the income effect can be positive or negative, a price change can have a small or a large effect on quantity demanded. In the unusual case of a so-called Giffen good, the quantity demanded may move in the same direction as the price change, thereby generating an upward-sloping individual demand curve.
- The market demand curve is the horizontal summation of the individual demand curves of all consumers in the market for a good. It can be used to calculate how much people value the consumption of particular goods and services.
- Demand is price inelastic when a 1-percent increase in price leads to a less than 1-percent decrease in quantity demanded, thereby increasing the consumer's expenditure. Demand is price elastic when a 1-percent increase in price leads to a more than 1-percent decrease in quantity demanded, thereby decreasing

the consumer's expenditure. Demand is unit elastic when a 1-percent increase in price leads to a 1-percent decrease in quantity demanded.

7. The concept of *consumer surplus* can be useful in determining the benefits that people receive from the consumption of a product. Consumer surplus is the difference between the maximum amount a consumer is willing to pay for a good and what he actually pays when buying it.
8. A network externality occurs when one person's demand is affected directly by the purchasing decisions of other consumers. A positive network externality, the bandwagon effect, occurs when a typical consumer's quantity demanded increases because she

considers it stylish to buy a product that others have purchased. Conversely, a negative network externality, the snob effect, occurs when the quantity demanded increases when fewer people own the good.

9. A number of methods can be used to obtain information about consumer demand. These include interview and experimental approaches, direct marketing experiments, and the more indirect statistical approach. The statistical approach can be very powerful in its application, but it is necessary to determine the appropriate variables that affect demand before the statistical work is done.

QUESTIONS FOR REVIEW

1. How is an individual demand curve different from a market demand curve? Which curve is likely to be more price elastic? (*Hint:* Assume that there are no network externalities.)
2. Is the demand for a particular brand of product, such as Head skis, likely to be more price elastic or price inelastic than the demand for the aggregate of all brands of downhill skis? Explain.
3. Tickets to a rock concert sell for \$10. At that price, however, the demand is substantially greater than the available number of tickets. Is the value or marginal benefit of an additional ticket greater than, less than, or equal to \$10? How might you determine that value?
4. Suppose a person allocates a given budget between two goods, food and clothing. If food is an inferior good, can you tell whether clothing is inferior or normal? Explain.
5. Which of the following combinations of goods are complements and which are substitutes? Could any of them be either in different circumstances? Discuss.
 - a. a mathematics class and an economics class
 - b. tennis balls and a tennis racket
 - c. steak and lobster
 - d. a plane trip and a train trip to the same destination
 - e. bacon and eggs
6. Which of the following events would cause a movement along the demand curve for U.S.-produced clothing? Which would cause a shift in the demand curve?
 - a. the removal of quotas on the importation of foreign clothes
 - b. an increase in the income of U.S. citizens
 - c. a cut in the industry's costs of producing domestic clothes that is passed on to the market in the form of lower prices
7. For which of the following goods is a price increase likely to lead to a substantial income (as well as substitution) effect?
 - a. salt
 - b. housing
 - c. theater tickets
 - d. food
8. Suppose that the average household in a state consumes 500 gallons of gasoline per year. A 10-cent gasoline tax is introduced, coupled with a \$50 annual tax rebate per household. Will the household be better or worse off after the new program is introduced?
9. Which of the following three groups is likely to have the most and which the least price-elastic demand for membership in the Association of Business Economists?
 - a. students
 - b. junior executives
 - c. senior executives

EXERCISES

1. The ACME Corporation determines that at current prices, the demand for its computer chips has a price elasticity of -2 in the short run. The price elasticity for its disc drives is -1 .
 - a. If ACME decides to raise the price of both products by 10 percent, what will happen to its sales? To its sales revenue?

- b. Can you tell from the available information which product will generate more revenue? If yes, which one? If not, what additional information would you need?
2. Refer to Example 4.3 on the aggregate demand for wheat in 1998. Consider 1996, at which time the domestic demand curve was $Q_{DD} = 1560 - 60P$. The export demand curve, however, was about the same as in 1998, i.e., $Q_{DE} = 1544 - 176P$. Calculate and draw the aggregate demand curve for wheat in 1996.
3. Judy has decided to allocate exactly \$500 to textbooks at college every year, even though she knows that the prices are likely to increase by from 5 to 10 percent per year and that she will be getting a substantial monetary gift from her grandparents next year. What is Judy's price elasticity of demand for textbooks? What is her income elasticity?
4. Vera has decided to upgrade the operating system on her new PC. She hears that the new Linux operating system is technologically superior to the Windows operating system and is substantially lower in price. However, when she asks her friends, it turns out they all use PCs with Windows. They agree that Linux is more appealing but add that they see relatively few copies of Linux on sale at local retail software stores. Based on what she learns and observes, Vera chooses to upgrade her PC with Windows. Can you explain her decision?
5. Suppose you are in charge of a toll bridge that is essentially cost free. The demand for bridge crossings Q is given by $P = 12 - 2Q$.
 - a. Draw the demand curve for bridge crossings.
 - b. How many people would cross the bridge if there were no toll?
 - c. What is the loss of consumer surplus associated with the charge of a \$6 toll?
6.
 - a. Orange juice and apple juice are known to be perfect substitutes. Draw the appropriate price-consumption curve (for a variable price of orange juice) and income-consumption curve.
 - b. Left shoes and right shoes are perfect complements. Draw the appropriate price-consumption and income-consumption curves.
7. Heather's marginal rate of substitution of movie tickets for video rentals is the same no matter how many videos she wants. Draw Heather's income consumption curve and her Engel curve for videos.
8. You are managing a \$300,000 city budget in which monies are spent on schools and public safety only. You are about to receive aid from the federal government to support a special antidrug program. Two programs are available: (1) a \$100,000 grant that must be spent on law enforcement; and (2) a 100-percent matching grant, in which each dollar of local spending on law enforcement is matched by a dollar of federal money. The federal matching program limits payment to each city to a maximum of \$100,000.

- a. Complete the table below with the amounts available for safety.

SCHOOLS	SAFETY (NO GOVT. ASSISTANCE)	SAFETY (PROGRAM 1)	SAFETY (PROGRAM 2)
\$0			
50,000			
100,000			
150,000			
200,000			
250,000			
300,000			

- b. Suppose that you allocate \$50,000 of the \$300,000 to schools. Which program would you (the manager) choose if you wish to maximize citizen satisfaction? What if you allocate \$250,000?
- c. Draw the budget constraints for the three options: no aid, program 1, or program 2.
9. By observing an individual's behavior in the situations outlined below, determine the relevant income elasticities of demand for each good (i.e., whether the good is normal or inferior). If you cannot determine the income elasticity, what additional information might you need?
 - a. Bill spends all his income on books and coffee. He finds \$20 while rummaging through a used paperback bin at the bookstore. He immediately buys a new hardcover book of poetry.
 - b. Bill loses \$10 with which he was going to buy a double espresso. He decides to sell his new book at a discount and use the money to buy coffee.
 - c. Being bohemian becomes the latest teen fad. As a result, coffee and book prices rise by 25 percent. Bill lowers his consumption of both goods by the same percentage.
 - d. Bill drops out of art school and gets an M.B.A. instead. He stops reading books and drinking coffee. Now he reads the *Wall Street Journal* and drinks bottled mineral water.
10. Suppose the income elasticity of demand for food is 0.5 and the price elasticity of demand -1.0 . Suppose also that Felicia spends \$10,000 a year on food, that the price of food is \$2, and that her income is \$25,000.
 - a. If a \$2 sales tax on food were to cause the price of food to double, what would happen to Felicia's consumption of food? (*Hint:* Because a large price change is involved, you should assume that the price elasticity measures an arc elasticity rather than a point elasticity.)

- b. Suppose that she is given a tax rebate of \$5000 to ease the effect of the tax. What would her consumption of food be now?
- c. Is she better or worse off when given a rebate equal to the sales tax payments? Discuss.
11. Suppose that you are the consultant to an agricultural cooperative that is deciding whether members should cut their production of cotton in half next year. The cooperative wants your advice as to whether this will increase the farmers' revenues. Knowing that cotton

(*c*) and watermelons (*w*) both compete for agricultural land in the South, you estimate the demand for cotton to be

$$c = 3.5 - 1.0 p_c + .25 p_w + .50 i$$

where p_c is the price of cotton, p_w the price of watermelon, and i income. Should you support or oppose the plan? Is there any additional information that would help you to provide a definitive answer?

APPENDIX TO CHAPTER 4

Demand Theory—A Mathematical Treatment

This appendix presents a mathematical treatment of the basics of demand theory. Our goal is to provide a short overview of the theory of demand for students who have some familiarity with the use of calculus. To do this, we will explain and then apply the concept of constrained optimization.

Utility Maximization

The theory of consumer behavior is based on the assumption that consumers maximize utility subject to the constraint of a limited budget. We saw in Chapter 3 that for each consumer, we can define a *utility function* that attaches a level of utility to each market basket. We also saw that the *marginal utility* of a good is defined as the change in utility associated with a one-unit increase in the consumption of the good. Using calculus, as we do in this appendix, marginal utility is measured as the utility change that results from a very small increase in consumption.

Suppose, for example, that Bob's utility function is given by $U(X, Y) = \log X + \log Y$, where, for the sake of generality, X is now used to represent food and Y represents clothing. In that case, the marginal utility associated with the additional consumption of X is given by the *partial derivative of the utility function with respect to good X* . Here, MU_X , representing the marginal utility of good X , is given by

$$\frac{\partial U(X, Y)}{\partial X} = \frac{\partial(\log X + \log Y)}{\partial X} = \frac{1}{X}$$

In the following analysis, we will assume, as in Chapter 3, that while the level of utility is an *increasing* function of the quantities of goods consumed, marginal utility *decreases* with consumption. When there are two goods, X and Y , the consumer's optimization problem may thus be written as

$$\text{Maximize } U(X, Y) \quad (\text{A4.1})$$

subject to the constraint that all income is spent on the two goods:

$$P_X X + P_Y Y = I \quad (\text{A4.2})$$

Here, $U(\)$ is the utility function, X and Y the quantities of the two goods purchased, P_X and P_Y the prices of the goods, and I income.¹

To determine the individual consumer's demand for the two goods, we choose those values of X and Y that maximize (A4.1) subject to (A4.2). When we know

In §3.1, we explain that a utility function is a formula that assigns a level of utility to each market basket.

In §3.2, marginal utility is described as the additional satisfaction obtained by consuming an additional amount of a good.

¹ To simplify the mathematics, we assume that the utility function is continuous (with continuous derivatives) and that goods are infinitely divisible.

the particular form of the utility function, we can solve to find the consumer's demand for X and Y directly. However, even if we write the utility function in its general form $U(X,Y)$, the technique of *constrained optimization* can be used to describe the conditions that must hold if the consumer is maximizing utility.

The Method of Lagrange Multipliers

The **method of Lagrange multipliers** is a technique that can be used to maximize or minimize a function subject to one or more constraints. Because we will use this technique to analyze production and cost issues later in the book, we will provide a step-by-step application of the method to the problem of finding the consumer's optimization given by equations (A4.1) and (A4.2).

method of Lagrange multipliers Technique to maximize or minimize a function subject to one or more constraints.

1. Stating the Problem First, we write the Lagrangian for the problem. The **Lagrangian** is the function to be maximized or minimized (here, utility is being maximized), plus a variable which we call λ times the constraint (here, the consumer's budget constraint). We will interpret the meaning of λ in a moment. The Lagrangian is then

$$\Phi = U(X,Y) - \lambda(P_X X + P_Y Y - I) \quad (\text{A4.3})$$

Note that we have written the budget constraint as

$$P_X X + P_Y Y - I = 0$$

i.e., as a sum of terms equal to zero. We then insert this sum into the Lagrangian.

2. Differentiating the Lagrangian If we choose values of X and Y that satisfy the budget constraint, then the second term in equation (A4.3) will be zero. Maximizing will therefore be equivalent to maximizing $U(X,Y)$. By differentiating Φ with respect to X, Y, and λ and then equating the derivatives to zero, we can obtain the necessary conditions for a maximum.² The resulting equations are

$$\begin{aligned} \frac{\partial \Phi}{\partial X} &= MU_X(X,Y) - \lambda P_X = 0 \\ \frac{\partial \Phi}{\partial Y} &= MU_Y(X,Y) - \lambda P_Y = 0 \\ \frac{\partial \Phi}{\partial \lambda} &= P_X X + P_Y Y - I = 0 \end{aligned} \quad (\text{A4.4})$$

Here as before, MU is short for *marginal utility*: In other words, $MU_X(X,Y) = \partial U(X,Y)/\partial X$, the change in utility from a very small increase in the consumption of good X.

² These conditions are necessary for an "interior" solution in which the consumer consumes positive amounts of both goods. The solution, however, could be a "corner" solution in which all of one good and none of the other is consumed.

3. Solving the Resulting Equations The three equations in (A4.4) can be rewritten as

$$\begin{aligned} MU_X &= \lambda P_X \\ MU_Y &= \lambda P_Y \\ P_X X + P_Y Y &= I \end{aligned}$$

Now we can solve these three equations for the three unknowns. The resulting values of X and Y are the solution to the consumer's optimization problem: They are the utility-maximizing quantities.

The Equal Marginal Principle

The third equation above is the consumer's budget constraint with which we started. The first two equations tell us that each good will be consumed up to the point at which the marginal utility from consumption is a multiple (λ) of the price of the good. To see the implication of this, we combine the first two conditions to obtain the *equal marginal principle*:

$$\lambda = \frac{MU_X(X,Y)}{P_X} = \frac{MU_Y(X,Y)}{P_Y} \quad (\text{A4.5})$$

In other words, the marginal utility of each good divided by its price is the same. To be optimizing, *the consumer must be getting the same utility from the last dollar spent by consuming either X or Y*. If this were not the case, consuming more of one good and less of the other would increase utility.

To characterize the individual's optimum in more detail, we can rewrite the information in (A4.5) to obtain

$$\frac{MU_X(X,Y)}{MU_Y(X,Y)} = \frac{P_X}{P_Y} \quad (\text{A4.6})$$

In other words, *the ratio of the marginal utilities is equal to the ratio of the prices*.

Marginal Rate of Substitution

We can use equation (A4.6) to see the link between utility functions and indifference curves that was spelled out in Chapter 3. An indifference curve represents all market baskets that give the consumer the same level of utility. If U^* is a fixed utility level, the indifference curve that corresponds to that utility level is given by

$$U(X,Y) = U^*$$

As the market baskets are changed by adding small amounts of X and subtracting small amounts of Y, the total change in utility must equal zero. Therefore

$$MU_X(X,Y)dX + MU_Y(X,Y)dY = dU^* = 0 \quad (\text{A4.7})$$

In §3.3, we show that the marginal rate of substitution is equal to the ratio of the marginal utilities of the two goods being consumed.

Rearranging,

$$-dY/dX = MU_X(X,Y)/MU_Y(X,Y) = MRS_{XY} \quad (\text{A4.8})$$

where MRS_{XY} represents the individual's marginal rate of substitution of X for Y . Because the left-hand side of (A4.8) represents the negative of the slope of the indifference curve, it follows that at the point of tangency, the individual's marginal rate of substitution (which trades off goods while keeping utility constant) is equal to the individual's ratio of marginal utilities, which in turn is equal to the ratio of the prices of the two goods, from (A4.6).³

When the individual indifference curves are convex, the tangency of the indifference curve to the budget line solves the consumer's optimization problem. This principle was illustrated by Figure 3.11 in Chapter 3.

Marginal Utility of Income

Whatever the form of the utility function, the Lagrange multiplier λ represents the extra utility generated when the budget constraint is relaxed—in this case by adding one dollar to the budget. To show how the principle works, we differentiate the utility function $U(X,Y)$ totally with respect to I :

$$dU/dI = MU_X(X,Y)(dX/dI) + MU_Y(X,Y)(dY/dI) \quad (\text{A4.9})$$

Because any increment in income must be divided between the two goods, it follows that

$$dI = P_X dX + P_Y dY \quad (\text{A4.10})$$

Substituting from (A4.5) into (A4.9), we get

$$dU/dI = \lambda P_X (dX/dI) + \lambda P_Y (dY/dI) = \lambda (P_X dX + P_Y dY)/dI \quad (\text{A4.11})$$

and substituting (A4.10) into (A4.11), we get

$$dU/dI = \lambda (P_X dX + P_Y dY)/(P_X dX + P_Y dY) = \lambda \quad (\text{A4.12})$$

Thus the Lagrange multiplier is the extra utility that results from an extra dollar of income.

Going back to our original analysis of the conditions for utility maximization, we see from equation (A4.5) that maximization requires that the utility obtained from the consumption of every good, per dollar spent on that good, be equal to the marginal utility of an additional dollar of income. If this were not the case, utility could be increased by spending more on the good with the higher ratio of marginal utility to price and less on the other good.

³ We implicitly assume that the "second-order conditions" for a utility maximum hold. The consumer, therefore, is maximizing rather than minimizing utility. The convexity condition is sufficient for the second-order conditions to be satisfied. In mathematical terms, the condition is that $d(MRS)/dX < 0$ or that $dY^2/dX^2 > 0$ where $-dY/dX$ is the slope of the indifference curve. Remember: diminishing marginal utility is not sufficient to ensure that indifference curves are convex.

An Example

In general, the three equations in (A4.4) can be solved to determine the three unknowns X , Y , and λ as a function of the two prices and income. Substitution for λ then allows us to solve for the demands for each of the two goods in terms of income and the prices of the two commodities. This principle can be most easily seen in terms of an example.

A frequently used utility function is the **Cobb-Douglas utility function**, which can be represented in two forms:

$$U(X,Y) = a \log(X) + (1 - a) \log(Y)$$

and

$$U(X,Y) = X^a Y^{1-a}$$

These two forms are equivalent for the purposes of demand theory because they both yield the identical demand functions for goods X and Y . We will derive the demand functions for the first form and leave the second as an exercise for the student.

To find the demand functions for X and Y , given the usual budget constraint, we first write the Lagrangian:

$$\Phi = a \log(X) + (1 - a) \log(Y) - \lambda (P_X X + P_Y Y - I)$$

Now differentiating with respect to X , Y , and λ and setting the derivatives equal to zero, we obtain

$$\partial \Phi / \partial X = a/X - \lambda P_X = 0$$

$$\partial \Phi / \partial Y = (1 - a)/Y - \lambda P_Y = 0$$

$$\partial \Phi / \partial \lambda = P_X X + P_Y Y - I = 0$$

The first two conditions imply that

$$P_X X = a/\lambda \quad (\text{A4.13})$$

$$P_Y Y = (1 - a)/\lambda \quad (\text{A4.14})$$

Combining these expressions with the last condition (the budget constraint) gives us

$$a/\lambda + (1 - a)/\lambda - I = 0$$

or $\lambda = 1/I$. Now we can substitute this expression for λ back into (A4.13) and (A4.14) to obtain the demand functions:

$$X = (a/P_X)I$$

$$Y = [(1 - a)/P_Y]I$$

Cobb-Douglas utility function
Utility function $U(X,Y) = X^a Y^{1-a}$,
where X and Y are two goods
and a is a constant.

In §2.3, we explain that the cross-price elasticity of demand refers to the percentage change in the quantity demanded of one good that results from a 1-percent increase in the price of another good.

In this example, the demand for each good depends only on the price of that good and on income, not on the price of the other good. Thus, the cross-price elasticities of demand are 0.

We can also use this example to review the meaning of Lagrange multipliers. To do so, let's substitute specific values for each of the parameters in the problem. Let $a = 1/2$, $P_X = \$1$, $P_Y = \$2$, and $I = \$100$. In this case, the choices that maximize utility are $X = 50$ and $Y = 25$. Also note that $\lambda = 1/100$. The Lagrange multiplier tells us that if an additional dollar of income were available to the consumer, the level of utility achieved would increase by $1/100$. This conclusion is relatively easy to check. With an income of \$101, the maximizing choices of the two goods are $X = 50.5$ and $Y = 25.25$. A bit of arithmetic tells us that the original level of utility is 3.565 and the new level of utility 3.575. As we can see, the additional dollar of income has indeed increased utility by .01, or $1/100$.

Duality in Consumer Theory

There are two different ways of looking at the consumer's optimization decision. The optimum choice of X and Y can be analyzed not only as the problem of choosing the highest indifference curve—the maximum value of $U(\)$ —that touches the budget line, but also as the problem of choosing the lowest budget line—the minimum budget expenditure—that touches a given indifference curve. We use the term **duality** to refer to these two perspectives. To see how this principle works, consider the following dual consumer optimization problem: the problem of minimizing the cost of achieving a particular level of utility:

$$\text{Minimize } P_X X + P_Y Y$$

subject to the constraint that

$$U(X, Y) = U^*$$

The corresponding Lagrangian is given by

$$\Phi = P_X X + P_Y Y - \mu(U(X, Y) - U^*) \quad (\text{A4.15})$$

where μ is the Lagrange multiplier. Differentiating Φ with respect to X , Y , and μ and setting the derivatives equal to zero, we find the following necessary conditions for expenditure minimization:

$$P_X - \mu MU_X(X, Y) = 0$$

$$P_Y - \mu MU_Y(X, Y) = 0$$

and

$$U(X, Y) = U^*$$

By solving the first two equations, we see that

$$\mu = [P_X / MU_X(X, Y)] = [P_Y / MU_Y(X, Y)] = 1/\lambda$$

Because it is also true that

$$MU_X(X, Y) / MU_Y(X, Y) = MRS_{XY} = P_X / P_Y$$

the cost-minimizing choice of X and Y must occur at the point of tangency of the budget line and the indifference curve that generates utility U^* . Because this is

duality Alternative way of looking at the consumer's utility maximization decision: Rather than choosing the highest indifference curve, given a budget constraint, the consumer chooses the lowest budget line that touches a given indifference curve.

the same point that maximized utility in our original problem, the dual expenditure minimization problem yields the same demand functions that are obtained from the direct utility maximization problem.

To see how the dual approach works, let's reconsider our Cobb-Douglas example. The algebra is somewhat easier to follow if we use the exponential form of the Cobb-Douglas utility function, $U(X, Y) = X^a Y^{1-a}$. In this case, the Lagrangian is given by

$$\Phi = P_X X + P_Y Y - \mu[X^a Y^{1-a} - U^*] \quad (\text{A4.16})$$

Differentiating with respect to X , Y , and μ and equating to zero, we obtain

$$P_X = \mu a U^* / X$$

$$P_Y = \mu(1 - a) U^* / Y$$

Multiplying the first equation by X and the second by Y and adding, we get

$$P_X X + P_Y Y = \mu U^*$$

First, we let I be the cost-minimizing expenditure (if the individual does not spend all of his income to get utility level U^* , U^* would not have maximized utility in the original problem). Then it follows that $\mu = I/U^*$. Substituting in the equations above, we obtain

$$X = aI/P_X \quad \text{and} \quad Y = (1 - a)I/P_Y$$

These are the same demand functions that we obtained before.

Income and Substitution Effects

The demand function tells us how any individual's utility-maximizing choices respond to changes in both income and the prices of goods. It is important, however, to distinguish that portion of any price change that involves *movement along an indifference curve* from that portion which involves *movement to a different indifference curve* (and therefore a change in purchasing power). To make this distinction, we consider what happens to the demand for good X when the price of X changes. As explained in Section 4.2, the change in demand can be divided into a *substitution effect* (the change in quantity demanded when the level of utility is fixed) and an *income effect* (the change in the quantity demanded with the level of utility changing but the relative price of good X unchanged). We denote the change in X that results from a unit change in the price of X , holding utility constant, by

$$\partial X / \partial P_X |_{U=U^*}$$

Thus the total change in the quantity demanded of X resulting from a unit change in P_X is

$$dX/dP_X = \partial X / \partial P_X |_{U=U^*} + (\partial X / \partial I)(\partial I / \partial P_X) \quad (\text{A4.17})$$

The first term on the right side of equation (A4.17) is the substitution effect (because utility is fixed); the second term is the income effect (because income increases).

From the consumer's budget constraint, $I = P_X X + P_Y Y$, we know by differentiation that

$$\partial I / \partial P_X = X \quad (\text{A4.18})$$

In §4.2, the effect of a price change is divided into an income effect and a substitution effect.

Suppose for the moment that the consumer owned goods X and Y . In that case, equation (A4.18) would tell us that when the price of good X increases by $\$1$, the amount of income that the consumer can obtain by selling the good increases by $\$X$. In our theory of the consumer, however, the consumer does not own the good. As a result, equation (A4.18) tells us how much additional income the consumer would need in order to be as well off after the price change as he was before. For this reason, it is customary to write the income effect as negative (reflecting a loss of purchasing power) rather than as a positive. Equation (A4.17) then appears as follows:

$$dX/dP_X = \partial X/\partial P_X|_{u=u^*} - X(\partial X/\partial I) \quad (\text{A4.19})$$

Slutsky equation Formula for decomposing the effects of a price change into substitution and income effects.

In this new form, called the **Slutsky equation**, the first term represents the *substitution effect*: the change in demand for good X obtained by keeping utility fixed. The second term is the *income effect*: the change in purchasing power resulting from the price change times the change in demand resulting from a change in purchasing power.

An alternative way to decompose a price change into substitution and income effects, which is usually attributed to John Hicks, does not involve indifference curves. In Figure A4.1, the consumer initially chooses market basket A on budget line RS . Suppose that after the price of food falls (and the budget line moves to RT), we take away enough income so that the individual is no better off (and

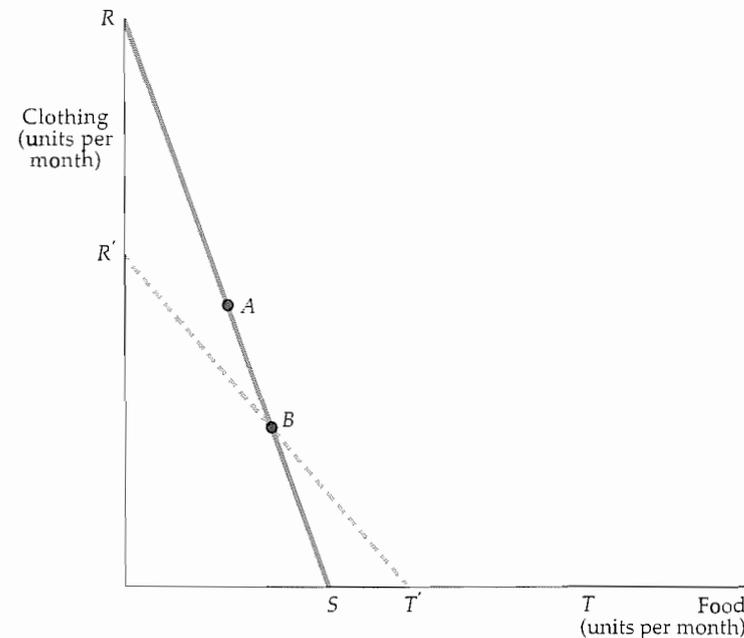


FIGURE A4.1 Hicksian Substitution Effect

The individual initially consumes market basket A . A decrease in the price of food shifts the budget line from RS to RT . If a sufficient amount of income is taken away from the individual to make him no better off than he was at A , two conditions must be met: The new market basket chosen must lie on line segment BT' of budget line $R'T'$ (which intersects RS to the right of A) and the quantity of food consumed must be greater than at A .

no worse off) than he was before. To do so, we draw a budget line parallel to RT . If the budget line passed through A , the consumer would be at least as satisfied as he was before the price change: He still has the option to purchase market basket A if he wishes. According to the **Hicksian substitution effect**, therefore, the budget line that leaves him equally well off must be a line such as $R'T'$, which is parallel to RT and which intersects RS at a point B below and to the right of point A .

Revealed preference tells us that the newly chosen market basket must lie on line segment BT' . Why? Because all market baskets on line segment $R'B$ could have been chosen but were not when the original budget line was RS . (Recall that the consumer preferred basket A to any other feasible market basket.) Now note that all points on line segment BT' involve more food consumption than does basket A . It follows that the quantity of food demanded increases whenever there is a decrease in the price of food with utility held constant. This negative substitution effect holds for all price changes and does not rely on the assumption of convexity of preferences that we made in Section 3.1.

Hicksian substitution effect Alternative to the Slutsky equation for decomposing price changes without recourse to indifference curves.

In §3.4, we explain how information about consumer preferences is revealed through the consumption choices that consumers make.

In §3.1, we explain that an indifference curve is convex if the marginal rate of substitution diminishes as we move down along the curve.

EXERCISES

- Which of the following utility functions are consistent with convex indifference curves and which are not?
 - $U(X,Y) = 2X + 5Y$
 - $U(X,Y) = (XY)^5$
 - $U(X,Y) = \text{Min}(X,Y)$, where Min is the minimum of the two values of X and Y
- Show that the two utility functions given below generate identical demand functions for goods X and Y :
 - $U(X,Y) = \log(X) + \log(Y)$
 - $U(X,Y) = (XY)^5$
- Assume that a utility function is given by $\text{Min}(X,Y)$, as in Exercise 1(c). What is the Slutsky equation that decomposes the change in the demand for X in response to a change in its price? What is the income effect? What is the substitution effect?

- Sharon has the following utility function:

$$U(X,Y) = \sqrt{X} + \sqrt{Y}$$

where X is her consumption of candy bars, with price $P_X = \$1$, and Y is her consumption of espressos, with $P_Y = \$3$.

- Derive Sharon's demand for candy bars and espresso.
- Assume that her income $I = \$100$. How many candy bars and how many espressos will Sharon consume?
- What is the marginal utility of income?

CHAPTER 5

Choice Under Uncertainty

So far, we have assumed that prices, incomes, and other variables are known with certainty. However, many of the choices that people make involve considerable uncertainty. Most people, for example, borrow to finance large purchases, such as a house or a college education, and plan to pay for them out of future income. But for most of us, future incomes are uncertain. Our earnings can go up or down; we can be promoted or demoted, or even lose our jobs. And if we delay buying a house or investing in a college education, we risk price rise increases that could make such purchases less affordable. How should we take these uncertainties into account when making major consumption or investment decisions?

Sometimes we must choose how much *risk* to bear. What, for example, should you do with your savings? Should you invest your money in something safe, such as a savings account, or something riskier but potentially more lucrative, such as the stock market? Another example is the choice of a job or career. Is it better to work for a large, stable company with job security but slim chance for advancement, or is it better to join (or form) a new venture that offers less job security but more opportunity for advancement?

To answer such questions, we must examine the ways that people can compare and choose among risky alternatives. We will do this by taking the following steps:

1. In order to compare the riskiness of alternative choices, we need to quantify risk. We therefore begin this chapter by discussing measures of risk.
2. We will examine people's preferences toward risk. Most people find risk undesirable, but some people find it more undesirable than others.
3. We will see how people can sometimes reduce or eliminate risk. Sometimes risk can be reduced by diversification, by buying insurance, or by investing in additional information.
4. In some situations, people must choose the amount of risk they wish to bear. A good example is investing in stocks or bonds. We will see that such investments involve trade-offs between the monetary gain that one can expect and the riskiness of that gain.

Chapter Outline

- 5.1 Describing Risk 150
- 5.2 Preferences Toward Risk 155
- 5.3 Reducing Risk 161
- *5.4 The Demand for Risky Assets 166

List of Examples

- 5.1 Deterring Crime 154
- 5.2 Business Executives and the Choice of Risk 160
- 5.3 The Value of Title Insurance When Buying a House 163
- 5.4 The Value of Information in the Dairy Industry 165
- 5.5 Investing in the Stock Market 173

5.1 Describing Risk

To describe risk quantitatively, we begin by listing all the possible outcomes of a particular action or event as well as the likelihood that each outcome will occur.¹ Suppose, for example, that you are considering investing in a company that explores for offshore oil. If the exploration effort is successful, the company's stock will increase from \$30 to \$40 per share; if not, the price will fall to \$20 per share. Thus there are two possible future outcomes: a \$40-per-share price and a \$20-per-share price.

Probability

probability Likelihood that a given outcome will occur.

Probability is the likelihood that a given outcome will occur. In our example, the probability that the oil exploration project is successful might be 1/4 and the probability that it is unsuccessful 3/4. (Note that the probabilities for all possible events must sum to 1.)

Our interpretation of probability can depend on the nature of the uncertain event, on the beliefs of the people involved, or both. One *objective* interpretation of probability relies on the frequency with which certain events tend to occur. Suppose we know that of the last 100 offshore oil explorations, 25 have succeeded and 75 failed. In that case, the probability of success of 1/4 is objective because it is based directly on the frequency of similar experiences.

But what if there are no similar past experiences to help measure probability? In such instances, objective measures of probability cannot be deduced and more subjective measures are needed. *Subjective probability* is the perception that an outcome will occur. This perception may be based on a person's judgment or experience, but not necessarily on the frequency with which a particular outcome has actually occurred in the past. When probabilities are subjectively determined, different people may attach different probabilities to different outcomes and thereby make different choices. For example, if the search for oil were to take place in an area where no previous searches had ever occurred, I might attach a higher subjective probability than you to the chance that the project will succeed: Perhaps I know more about the project or I have a better understanding of the oil business and can therefore make better use of our common information. Either different information or different abilities to process the same information can cause subjective probabilities to vary among individuals.

Regardless of the interpretation of probability, it is used in calculating two important measures that help us describe and compare risky choices. One measure tells us the *expected value* and the other the *variability* of the possible outcomes.

Expected Value

The **expected value** associated with an uncertain situation is a weighted average of the **payoffs** or values resulting from all possible outcomes. The probabilities of each outcome are used as weights. Thus the expected value measures the *central tendency*—that is, the payoff or value that we would expect on average.

¹ Some people distinguish between uncertainty and risk along the lines suggested some 60 years ago by economist Frank Knight. *Uncertainty* can refer to situations in which many outcomes are possible but their likelihoods unknown. *Risk* then refers to situations in which we can list all possible outcomes and know the likelihood of each occurring. In this chapter, we will always refer to risky situations but will simplify the discussion by using *uncertainty* and *risk* interchangeably.

Our offshore oil exploration example had two possible outcomes: Success yields a payoff of \$40 per share, failure a payoff of \$20 per share. Denoting "probability of" by Pr, we express the expected value in this case as

$$\begin{aligned} \text{Expected value} &= \text{Pr}(\text{success})(\$40/\text{share}) + \text{Pr}(\text{failure})(\$20/\text{share}) \\ &= (1/4)(\$40/\text{share}) + (3/4)(\$20/\text{share}) = \$25/\text{share} \end{aligned}$$

More generally, if there are two possible outcomes having payoffs X_1 and X_2 and if the probabilities of each outcome are given by Pr_1 and Pr_2 , then the expected value is

$$E(X) = \text{Pr}_1 X_1 + \text{Pr}_2 X_2$$

When there are n possible outcomes, the expected value becomes

$$E(X) = \text{Pr}_1 X_1 + \text{Pr}_2 X_2 + \dots + \text{Pr}_n X_n$$

Variability

Variability is the extent to which the possible outcomes of an uncertain situation differ. To see why variability is important, suppose you are choosing between two part-time sales jobs that have the same expected income (\$1500). The first job is based entirely on commission—the income earned depends on how much you sell. There are two equally likely payoffs for this job: \$2000 for a successful sales effort and \$1000 for one that is less successful. The second job is salaried. It is very likely (.99 probability) that you will earn \$1510, but there is a .01 probability that the company will go out of business, in which case you would earn \$510 in severance pay. Table 5.1 summarizes these possible outcomes, their payoffs, and their probabilities.

Note that these two jobs have the same expected income. For Job 1, expected income is $.5(\$2000) + .5(\$1000) = \$1500$; for Job 2 it is $.99(\$1510) + .01(\$510) = \$1500$. However, the *variability* of the possible payoffs is different. We measure variability by recognizing that large differences between actual and expected payoffs (whether positive or negative) imply greater risk. We call these differences **deviations**. Table 5.2 shows the deviations of the possible incomes from the expected income from each of the two jobs.

variability Extent to which possible outcomes of an uncertain event may differ.

deviation Difference between expected payoff and actual payoff.

TABLE 5.1 Income from Sales Jobs

	OUTCOME 1		OUTCOME 2		Expected Income (\$)
	Probability	Income (\$)	Probability	Income (\$)	
Job 1: Commission	.5	2000	.5	1000	1500
Job 2: Fixed salary	.99	1510	.01	510	1500

TABLE 5.2 Deviations from Expected Income (\$)

	OUTCOME 1	DEVIATION	OUTCOME 2	DEVIATION
Job 1	2000	500	1000	- 500
Job 2	1510	10	510	- 990

expected value Probability-weighted average of the values associated with all possible outcomes.

payoff Value associated with a possible outcome.

TABLE 5.3 Calculating Variance (\$)

	OUTCOME 1	DEVIATION SQUARED	OUTCOME 2	DEVIATION SQUARED	AVERAGE DEVIATION SQUARED	STANDARD DEVIATION
Job 1	2000	250,000	1000	250,000	250,000	500
Job 2	1510	100	510	980,100	9,900	99.50

By themselves, deviations do not provide a measure of variability. Why? Because they are sometimes positive and sometimes negative, and as you can see from Table 5.2, the average deviation is always 0.² To get around this problem, we square each deviation, yielding numbers that are always positive. We then measure variability by calculating the **standard deviation**: the square root of the average of the *squares* of the deviations of the payoffs associated with each outcome from their expected value.³

standard deviation Square root of the average of the squares of the deviations of the payoffs associated with each outcome from their expected values.

Table 5.3 shows the calculation of the standard deviation for our example. Note that the average of the squared deviations under Job 1 is given by

$$.5(\$250,000) + .5(\$250,000) = \$250,000$$

The standard deviation is therefore equal to the square root of \$250,000, or \$500. Likewise, the average of the squared deviations under Job 2 is given by

$$.99(\$100) + .01(\$980,100) = \$9900$$

The standard deviation is the square root of \$9,900, or \$99.50. Thus the second job is much less risky than the first; the standard deviation of the incomes is much lower.⁴

The concept of standard deviation applies equally well when there are many outcomes rather than just two. Suppose, for example, that the first job yields incomes ranging from \$1000 to \$2000 in increments of \$100 that are all equally likely. The second job yields incomes from \$1300 to \$1700 (again in increments of \$100) that are also equally likely. Figure 5.1 shows the alternatives graphically. (If there had been only two equally probable outcomes, then the figure would be drawn as two vertical lines, each with a height of 0.5.)

You can see from Figure 5.1 that the first job is riskier than the second. The “spread” of possible payoffs for the first job is much greater than the spread for the second. As a result, the standard deviation of the payoffs associated with the first job is greater than that associated with the second.

In this particular example, all payoffs are equally likely. Thus the curves describing the probabilities for each job are flat. In many cases, however, some

² For Job 1, the average deviation is $.5(\$500) + .5(-\$500) = 0$; for Job 2 it is $.99(\$10) + .01(-\$990) = 0$.

³ Another measure of variability, *variance*, is the square of the standard deviation.

⁴ In general, when there are two outcomes with payoffs X_1 and X_2 , occurring with probability Pr_1 and Pr_2 , and $E(X)$ is the expected value of the outcomes, the standard deviation is given by σ , where

$$\sigma^2 = Pr_1[(X_1 - E(X))^2] + Pr_2[(X_2 - E(X))^2]$$

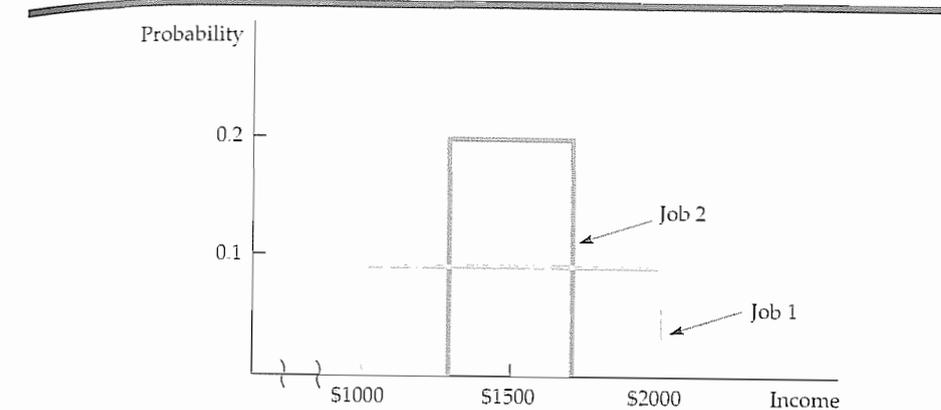


FIGURE 5.1 Outcome Probabilities for Two Jobs

The distribution of payoffs associated with Job 1 has a greater spread and a greater standard deviation than the distribution of payoffs associated with Job 2. Both distributions are flat because all outcomes are equally likely.

payoffs are more likely than others. Figure 5.2 shows a situation in which the most extreme payoffs are the least likely. Again, the salary from Job 1 has a greater standard deviation. From this point on, we will use the standard deviation of payoffs to measure degree of risk.

Decision Making

Suppose you are choosing between the two sales jobs described in our original example. Which job would you take? If you dislike risk, you will take the second job: It offers the same expected income as the first but with less risk. But suppose

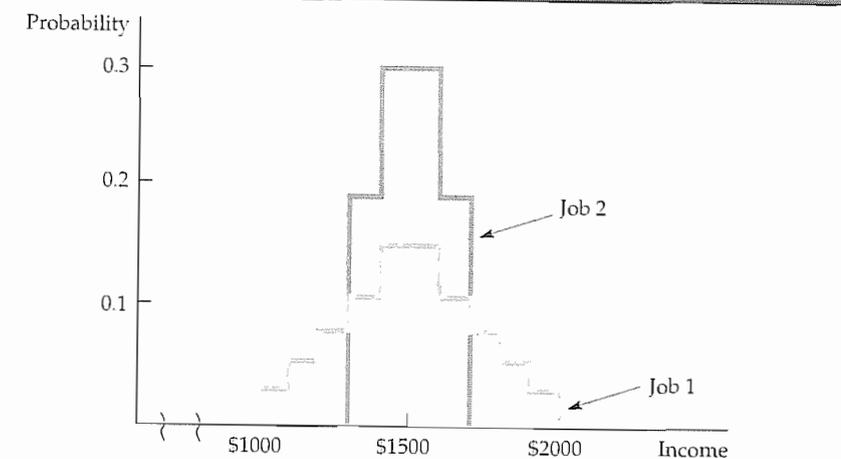


FIGURE 5.2 Unequal Probability Outcomes

The distribution of payoffs associated with Job 1 has a greater spread and a greater standard deviation than the distribution of payoffs associated with Job 2. Both distributions are peaked because the extreme payoffs are less likely than those near the middle of the distribution.

TABLE 5.4 Incomes from Sales Jobs—Modified (\$)

	OUTCOME 1	DEVIATION SQUARED	OUTCOME 2	DEVIATION SQUARED	EXPECTED INCOME	STANDARD DEVIATION
Job 1	2,100	250,000	1,100	250,000	1,600	500
Job 2	1,510	100	510	980,100	1,500	99.50

we add \$100 to each of the payoffs in the first job, so that the expected payoff increases from \$1500 to \$1600. Table 5.4 gives the new earnings and the squared deviations.

The two jobs can now be described as follows:

Job 1:	Expected income = \$1600	Standard deviation = \$500
Job 2:	Expected income = \$1500	Standard deviation = \$99.50

Job 1 offers a higher expected income but is much riskier than Job 2. Which job is preferred depends on the individual. While an aggressive entrepreneur who doesn't mind taking risks might choose Job 1, with the higher expected income and higher standard deviation, a more conservative person might choose the second job.

People's attitudes toward risk affect many of the decisions they make. In Example 5.1 we will see how attitudes toward risk affect people's willingness to break the law, and how this has implications for the fines that should be set for various violations. Then in Section 5.2, we will further develop our theory of consumer choice by examining people's risk preferences in greater detail.

EXAMPLE 5.1 Detering Crime

Fines may be better than incarceration in deterring certain types of crimes, such as speeding, double-parking, tax evasion, and air polluting.⁵ A person choosing to violate the law in these ways has good information and can reasonably be assumed to be behaving rationally.

Other things being equal, the greater the fine, the more a potential criminal will be discouraged from committing the crime. For example, if it cost nothing to catch criminals and if the crime imposed a calculable cost of \$1000 on society, we might choose to catch all violators and impose a fine of \$1000 on each. This practice would discourage people whose benefit from engaging in the activity was less than the \$1000 fine.

In practice, however, it is very costly to catch lawbreakers. Therefore, we save on administrative costs by imposing relatively high fines (which are no more costly to collect than low fines), while allocating resources so that only a

⁵ This discussion builds indirectly on Gary S. Becker, "Crime and Punishment: An Economic Approach," *Journal of Political Economy* (March/April 1968): 169–217. See also Mitchell Polinsky and Steven Shavell, "The Optimal Tradeoff Between the Probability and the Magnitude of Fines," *American Economic Review* 69 (December 1979): 880–91.

fraction of the violators are apprehended. Thus the size of the fine that must be imposed to discourage criminal behavior depends on the attitudes toward risk of potential violators.

Suppose that a city wants to deter people from double-parking. By double-parking, a typical resident saves \$5 in terms of his own time for engaging in activities that are more pleasant than searching for a parking space. If it cost nothing to catch a double-parker, a fine of just over \$5—say, \$6—should be assessed every time he double-parked. This policy will ensure that the net benefit of double-parking (the \$5 benefit less the \$6 fine) would be less than zero. He will therefore choose to obey the law. In fact, all potential violators whose benefit was less than or equal to \$5 would be discouraged, while a few whose benefit was greater than \$5 (say, someone who double-parks because of an emergency) would violate the law.

In practice, it is too costly to catch all violators. Fortunately, it's also unnecessary. The same deterrence effect can be obtained by assessing a fine of \$50 and catching only one in ten violators (or perhaps a fine of \$500 with a one-in-100 chance of being caught). In each case, the expected penalty is \$5, i.e., $[\$50][.1]$ or $[\$500][.01]$. A policy that combines a high fine and a low probability of apprehension is likely to reduce enforcement costs. This approach is especially effective if drivers don't like to take risks. In our example, a \$50 fine with a .1 probability of being caught might discourage most people from violating the law. We will examine attitudes toward risk in the next section.

5.2 Preferences Toward Risk

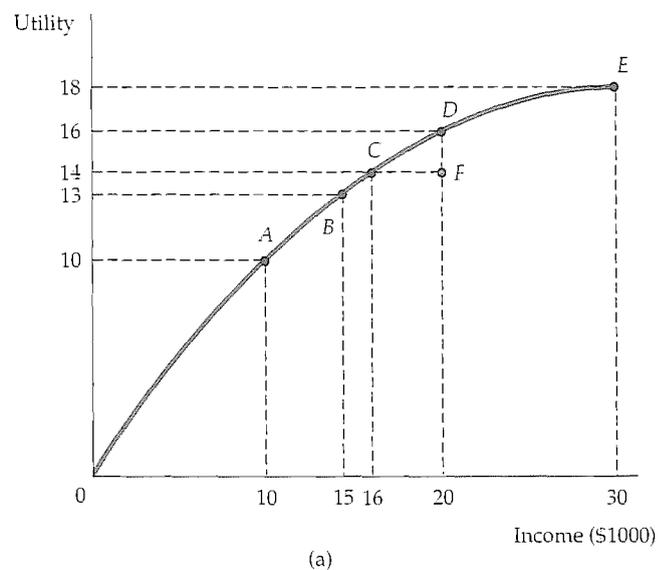
We used a job example to show how people might evaluate risky outcomes, but the principles apply equally well to other choices. In this section, we concentrate on consumer choices generally and on the *utility* that consumers obtain from choosing among risky alternatives. To simplify things, we'll consider the utility that a consumer gets from his or her income—or, more appropriately, the market basket that the consumer's income can buy. We now measure payoffs, therefore, in terms of utility rather than dollars.

Figure 5.3(a) shows how we can describe one woman's preferences toward risk. The curve OE , which gives her utility function, tells us the level of utility (on the vertical axis) that she can attain for each level of income (measured in thousands of dollars on the horizontal axis). The level of utility increases from 10 to 16 to 18 as income increases from \$10,000 to \$20,000 to \$30,000. But note that *marginal utility* is diminishing, falling from 10 when income increases from 0 to \$10,000, to 6 when income increases from \$10,000 to \$20,000, and to 2 when income increases from \$20,000 to \$30,000.

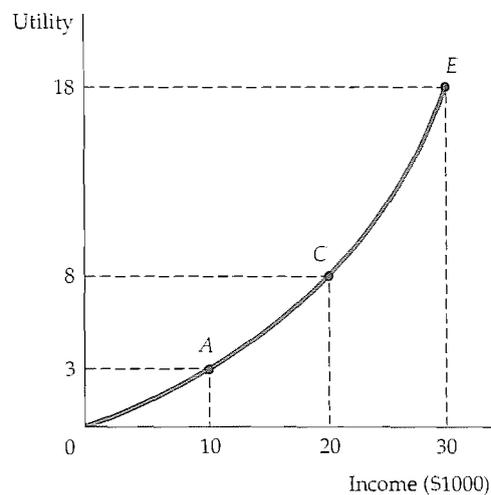
Now suppose that our consumer has an income of \$15,000 and is considering a new but risky sales job that will either double her income to \$30,000 or cause it to fall to \$10,000. Each possibility has a probability of .5. As Figure 5.3(a) shows, the utility level associated with an income of \$10,000 is 10 (at point A) and the utility level associated with an income of \$30,000 is 18 (at E). The risky job must be compared with the current \$15,000 job, for which the utility is 13 (at B).

In §3.1, we explained that a utility function assigns a level of utility to each possible market basket.

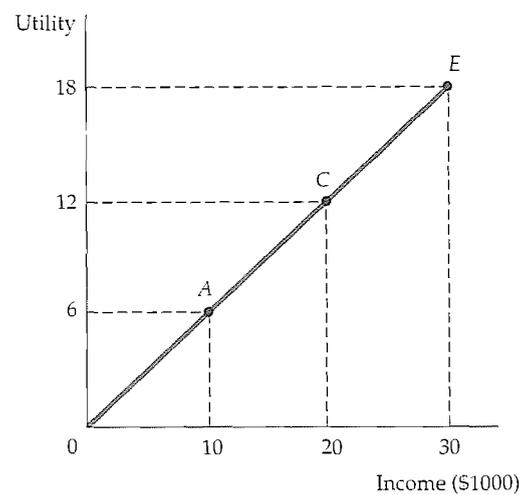
In §3.2, marginal utility is described as the additional satisfaction obtained by consuming an additional amount of a good.



(a)



(b)



(c)

FIGURE 5.3 Risk Aversion

People differ in their preferences toward risk. In (a), a consumer's marginal utility diminishes as income increases. The consumer is risk averse because she would prefer a certain income of \$20,000 (with a utility of 16) to a gamble with a .5 probability of \$10,000 and a .5 probability of \$30,000 (and expected utility of 14). In (b), the consumer is risk loving: She would prefer the same gamble (with expected utility of 10.5) to the certain income (with a utility of 8). Finally, the consumer in (c) is risk neutral and indifferent between certain events and uncertain events with the same expected income.

expected utility Sum of the utilities associated with all possible outcomes, weighted by the probability that each outcome will occur.

To evaluate the new job, she can calculate the expected value of the resulting income. Because we are measuring value in terms of the woman's utility, we must calculate the **expected utility** $E(u)$ that she can obtain. The expected utility is the sum of the utilities associated with all possible outcomes, weighted by the probability that each outcome will occur. In this case expected utility is

$$E(u) = (1/2)u(\$10,000) + (1/2)u(\$30,000) = 0.5(10) + 0.5(18) = 14$$

The new, risky job is thus preferred to the original job because the expected utility of 14 is greater than the original utility of 13.

The old job involved no risk—it guaranteed an income of \$15,000 and a utility level of 13. The new job is risky but offers both a higher expected income (\$20,000) and, more importantly, a higher expected utility. If the woman wishes to increase her expected utility, she will take the risky job.

Different Preferences Toward Risk

People differ in their willingness to bear risk. Some are risk averse, some risk loving, and some risk neutral. An individual who is **risk averse** prefers a certain given income to a risky income with the same expected value. (Such a person has a diminishing marginal utility of income.) Risk aversion is the most common attitude toward risk. To see that most people are risk averse most of the time, note that most people not only buy life insurance, health insurance, and car insurance, but also seek occupations with relatively stable wages.

risk averse Preferring a certain income to a risky income with the same expected value.

Figure 5.3(a) applies to a woman who is risk averse. Suppose she can have either a certain income of \$20,000, or a job yielding an income of \$30,000 with probability .5 and an income of \$10,000 with probability .5 (so that the expected income is \$20,000). As we saw, the expected utility of the uncertain income is 14—an average of the utility at point A (10) and the utility at E (18)—and is shown by F. Now we can compare the expected utility associated with the risky job to the utility generated if \$20,000 were earned without risk. This latter utility level, 16, is given by D in Figure 5.3(a). It is clearly greater than the expected utility of 14 associated with the risky job.

For a risk-averse person, losses are more important (in terms of the change in utility) than gains. Again, this can be seen from Figure 5.3(a). A \$10,000 increase in income, from \$20,000 to \$30,000, generates an increase in utility of two units; a \$10,000 decrease in income, from \$20,000 to \$10,000, creates a loss of utility of six units.

A person who is **risk neutral** is indifferent between a certain income and an uncertain income with the same expected value. In Figure 5.3(c) the utility associated with a job generating an income of either \$10,000 or \$30,000 with equal probability is 12, as is the utility of receiving a certain income of \$20,000. As you can see from the figure, the marginal utility of income is constant for a risk-neutral person.⁶

risk neutral Being indifferent between a certain income and an uncertain income with the same expected value.

Finally, an individual who is **risk loving** prefers an uncertain income to a certain one, even if the expected value of the uncertain income is less than that of the certain income. Figure 5.3(b) shows this third possibility. In this case, the expected utility of an uncertain income, which will be either \$10,000 with probability .5 or \$30,000 with probability .5, is higher than the utility associated with a certain income of \$20,000. Numerically,

risk loving Preferring a risky income to a certain income with the same expected value.

$$E(u) = .5u(\$10,000) + .5u(\$30,000) = .5(3) + .5(18) = 10.5 > u(\$20,000) = 8$$

Of course some people may be averse to some risks and act like risk lovers with respect to others. For example, many people purchase life insurance and are conservative with respect to their choice of jobs, but still enjoy gambling.

⁶ Thus when people are risk neutral, the income they earn can be used as an indicator of well-being. A government policy that doubles incomes would then also double their utility. At the same time, government policies that alter the risks that people face, without changing their expected incomes, would not affect their well-being. Risk neutrality allows a person to avoid the complications that might be associated with the effects of governmental actions on the riskiness of outcomes.

risk premium Maximum amount of money that a risk-averse person will pay to avoid taking a risk.

Some criminologists might describe criminals as risk lovers, especially if they commit crimes despite a high prospect of apprehension and punishment. Except for such special cases, however, few people are risk loving, at least with respect to major purchases or large amounts of income or wealth.

Risk Premium The risk premium is the maximum amount of money that a risk-averse person will pay to avoid taking a risk. In general, the magnitude of the risk premium depends on the risky alternatives that the person faces. To determine the risk premium, we have reproduced the utility function of Figure 5.3(a) in Figure 5.4 and extended it to an income of \$40,000. Recall that an expected utility of 14 is achieved by a woman who is going to take a risky job with an expected income of \$20,000. This outcome is shown graphically by drawing a horizontal line to the vertical axis from point *F*, which bisects straight line *AE* (thus representing an average of \$10,000 and \$30,000). But the utility level of 14 can also be achieved if the woman has a certain income of \$16,000, as shown by dropping a vertical line from point *C*. Thus the risk premium of \$4,000, given by line segment *CF*, is the amount of expected income (\$20,000 minus \$16,000) that she would give up in order to remain indifferent between the risky job and the safe one.

Risk Aversion and Income The extent of an individual's risk aversion depends on the nature of the risk and on the person's income. Other things being equal, risk-averse people prefer a smaller variability of outcomes. We saw that when there are two outcomes—an income of \$10,000 and an income of \$30,000—the risk premium is \$4,000. Now consider a second risky job, involving a .5 probability of receiving an income of \$40,000 and, as shown in Figure 5.4, with a

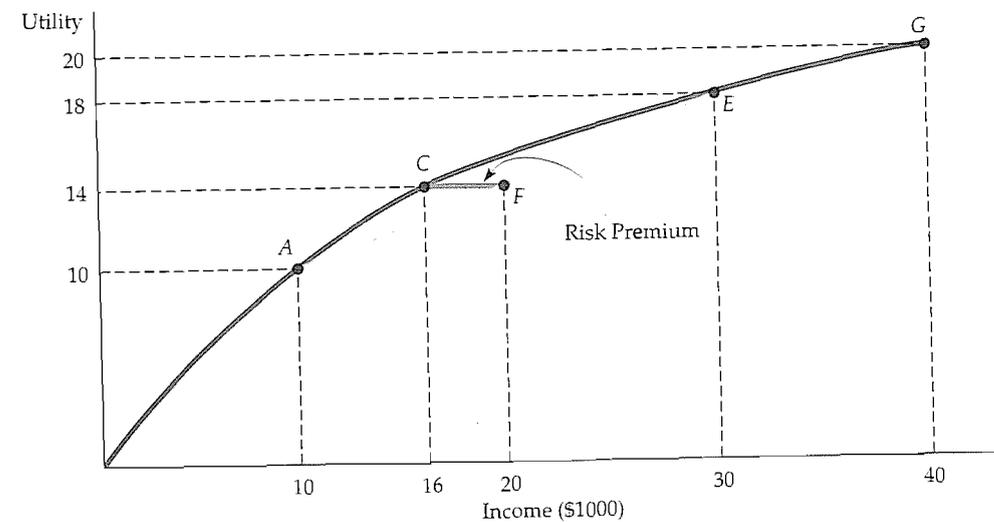


FIGURE 5.4 Risk Premium

The risk premium, *CF*, measures the amount of income that an individual would give up to leave her indifferent between a risky choice and a certain one. Here, the risk premium is \$4,000 because a certain income of \$16,000 (at point *C*) gives her the same expected utility (14) as the uncertain income (a .5 probability of being at point *A* and a .5 probability of being at point *E*) that has an expected value of \$20,000.

utility level of 20; and a .5 probability of getting an income of \$0, with a utility level of 0. The expected income is again \$20,000, but the expected utility is only 10:

$$\text{Expected utility} = .5u(\$0) + .5u(\$40,000) = 0 + .5(20) = 10$$

Because the utility of having a certain income of \$20,000 is 16, our consumer loses 6 units of utility if she is required to accept the job. The risk premium in this case is equal to \$10,000 because the utility of a certain income of \$10,000 is 10: She is willing to give up \$10,000 of her \$20,000 expected income to ensure a certain income of \$10,000 with the same level of expected utility. Thus the greater the variability, the more a person is willing to pay to avoid a risky situation.

Risk Aversion and Indifference Curves We can also describe the extent of a person's risk aversion in terms of indifference curves that relate expected income to the variability of income, where the latter is measured by the standard deviation. Figure 5.5 shows such indifference curves for two individuals, one who is very risk averse and another who is only slightly risk averse. Each indifference curve shows the combinations of expected income and standard deviation of income that give the individual the same amount of utility. Observe that all of the indifference curves are upward sloping: Because risk is undesirable, the greater the amount of risk, the greater the expected income needed to make the individual equally well off.

Figure 5.5(a) describes an individual who is highly risk averse. Observe that an increase in the standard deviation of income requires a large increase in expected income to leave this person equally well off. Figure 5.5(b) applies to a slightly risk-averse person. In this case, a large increase in the standard deviation of income requires only a small increase in expected income.

In §3.1, we define an indifference curve to be all market baskets that generate the same level of satisfaction for a consumer.

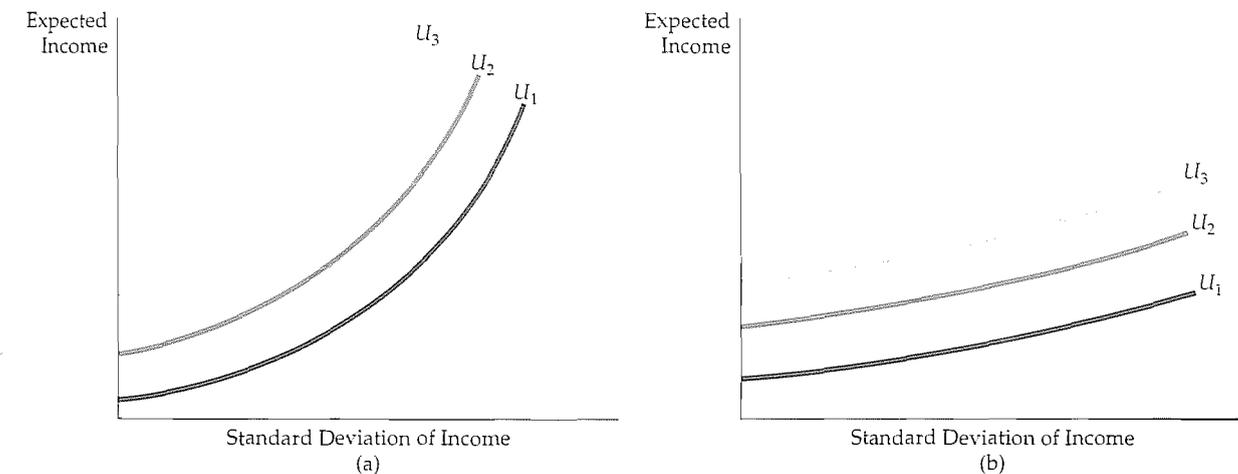


FIGURE 5.5 Risk Aversion and Indifference Curves

Part (a) applies to a person who is highly risk averse: An increase in this individual's standard deviation of income requires a large increase in expected income if he is to remain equally well off. Part (b) applies to a person who is only slightly risk averse: An increase in the standard deviation of income requires only a small increase in expected income if she is to remain equally well off.

We will return to the use of indifference curves as a means of describing risk aversion in Section 5.4, where we discuss the demand for risky assets. First, however, we will turn to the ways in which an individual can reduce risk.

EXAMPLE 5.2 Business Executives and the Choice of Risk

Are business executives more risk loving than most people? When they are presented with alternative strategies, some risky, some safe, which do they choose? In one study, 464 executives were asked to respond to a questionnaire describing risky situations that an individual might face as vice president of a hypothetical company.⁷ Respondents were presented with four risky events, each of which had a given probability of a favorable and unfavorable outcome. The payoffs and probabilities were chosen so that each event had the same expected value. In increasing order of the risk involved (as measured by the difference between the favorable and unfavorable outcomes), the four items were:

1. A lawsuit involving a patent violation
2. A customer threat concerning the supplying of a competitor
3. A union dispute
4. A joint venture with a competitor

To gauge their willingness to take or avoid risks, researchers asked respondents a series of questions. In different situations, they could opt to delay a choice, to collect information, to bargain, or to delegate a decision. Each option permitted respondents to avoid taking risks or to modify the risks that they would take later.

The study found that executives vary substantially in their preferences toward risk. Roughly 20 percent indicated that they were relatively neutral toward risk; 40 percent opted for the more risky alternatives; and 20 percent were clearly risk averse (20 percent did not respond). More importantly, executives (including those who chose risky alternatives) typically made efforts to reduce or eliminate risk, usually by delaying decisions and collecting more information.

In general, risk can arise when the expected gain is either positive (e.g., a chance for a large reward versus a small one) or negative (e.g., a chance for a large loss or for no loss). The study found that differing preferences toward risk depended on whether a given risk involved gains or losses. In general, those who liked risky situations did so when losses were involved. (Perhaps they were willing to gamble against a large loss in the hope of breaking even.) However, when the risks involved gains, the same executives were more conservative, opting for the less risky alternatives.⁸

⁷ This example is based on Kenneth R. MacCrimmon and Donald A. Wehrung, "The Risk In-Basket," *Journal of Business* 57 (1984): 367-87.

⁸ Interestingly, some people treat the risk of a small gain in income very differently from the risk of a small loss. *Prospect theory*, developed by psychologists Daniel Kahneman and Amos Tversky, helps to explain this phenomenon. See "Rational Choice and the Framing of Decisions," *Journal of Business* 59 (1986): S251-78, and "Prospect Theory: An Analysis of Decision under Risk," *Econometrica* 47 (1979): 263-92.

5.3 Reducing Risk

As the recent growth in state lotteries shows, people sometimes choose risky alternatives that suggest risk-loving rather than risk-averse behavior. In the face of a broad variety of risky situations, however, people are generally risk averse. In this section, we describe three ways by which consumers and managers commonly reduce risks: *diversification*, *insurance*, and *obtaining more information* about choices and payoffs.

Diversification

Recall the old saying, "Don't put all your eggs in one basket." Ignoring this advice is unnecessarily risky: If your basket turns out to be a bad bet, all will be lost. Instead, one can reduce risk through **diversification**: allocating one's resources to a variety of risky situations.

Suppose, for example, that you plan to take a part-time job selling appliances on a commission basis. You can decide to sell only air conditioners or only heaters, or you can spend half your time selling each. Of course, you can't be sure how hot or cold the weather will be next year. How should you apportion your time in order to minimize the risk involved in the job?

Risk can be minimized by *diversification*—by allocating your time so that you sell two or more products (whose sales are not closely related) rather than a single product. Suppose there is a 0.5 probability that it will be a relatively hot year, and a 0.5 probability that it will be cold. Table 5.5 gives the earnings that you can make selling air conditioners and heaters.

If you sell only air conditioners or only heaters, your actual income will be either \$12,000 or \$30,000, but your expected income will be \$21,000 ($.5[\$30,000] + .5[\$12,000]$). But suppose you diversify by dividing your time evenly between the two products. In that case, your income will certainly be \$21,000, regardless of the weather. If the weather is hot, you will earn \$15,000 from air conditioner sales and \$6,000 from heater sales; if it is cold, you will earn \$6,000 from air conditioners and \$15,000 from heaters. In this instance, diversification eliminates all risk.

Of course, diversification is not always this easy. In our example, heater and air conditioner sales are **negatively correlated**—they tend to move in opposite directions. In other words, whenever sales of one are strong, sales of the other are weak. But the principle of diversification is a general one: As long as you can allocate your resources toward a variety of activities whose outcomes are *not* closely related, you can eliminate some risk.

The Stock Market Diversification is especially important for people who invest in the stock market. On any given day, the price of an individual stock can go up or down by a large amount, but some stocks rise in price while others fall.

diversification Reducing risk by allocating resources to a variety of activities whose outcomes are not closely related.

negatively correlated Having a tendency to move in opposite directions (said of two variables).

TABLE 5.5 Income from Sales of Appliances (\$)

	HOT WEATHER	COLD WEATHER
Air conditioner sales	30,000	12,000
Heater sales	12,000	30,000

An individual who invests all her money in a single stock (i.e., puts all her eggs in one basket) is therefore taking much more risk than is necessary. Risk can be reduced—although not eliminated—by investing in a portfolio of ten or twenty different stocks. Equivalently, you can diversify by buying shares in *mutual funds*: organizations that pool funds of individual investors to buy a large number of different stocks.

In the case of the stock market, not all risk is diversifiable. Although some stocks go up in price when others go down, stock prices are to some extent **positively correlated**: they tend to move in the same direction in response to changes in economic conditions. For example, the onset of a severe recession, which is likely to reduce the profits of many companies, may be accompanied by a decline in the overall market. Even with a diversified portfolio of stocks, therefore, you still face some risk.

Insurance

We have seen that risk-averse people are willing to pay to avoid risk. In fact, if the cost of insurance is equal to the expected loss (e.g., a policy with an expected loss of \$1000 will cost \$1000), risk-averse people will buy enough insurance to recover fully from any financial losses they might suffer.

Why? The answer is implicit in our discussion of risk aversion. Buying insurance assures a person of having the same income whether or not there is a loss. Because the insurance cost is equal to the expected loss, this certain income is equal to the expected income from the risky situation. For a risk-averse consumer, the guarantee of the same income regardless of the outcome generates more utility than would be the case if that person had a high income when there was no loss and a low income when a loss occurred.

To clarify this point, let's suppose a homeowner faces a 10-percent probability that his house will be burglarized and he will suffer a \$10,000 loss. Let's assume he has \$50,000 worth of property. Table 5.6 shows his wealth in two situations—with insurance costing \$1000 and without insurance.

Note that expected wealth is the same (\$49,000) in both situations. The variability, however, is quite different: As the table shows, with no insurance the standard deviation of wealth is \$3000, whereas with insurance it is 0. If there is no burglary, the uninsured homeowner gains \$1000 relative to the insured homeowner. But with a burglary, the uninsured homeowner loses \$9000 relative to the insured homeowner. Remember: for a risk-averse individual, losses count more (in terms of changes in utility) than gains. A risk-averse homeowner, therefore, will enjoy higher utility by purchasing insurance.

The Law of Large Numbers Consumers usually buy insurance from companies that specialize in selling it. Insurance companies are firms that offer insurance because they know that when they sell a large number of policies,

positively correlated Having a tendency to move in the same direction.

they face relatively little risk. The ability to avoid risk by operating on a large scale is based on the *law of large numbers*, which tells us that although single events may be random and largely unpredictable, the average outcome of many similar events can be predicted. For example, I may not be able to predict whether a coin toss will come out heads or tails, but I know that when many coins are flipped, approximately half will turn up heads and half tails. Likewise, if I am selling automobile insurance, I cannot predict whether a particular driver will have an accident, but I can be reasonably sure, judging from past experience, about how many accidents a large group of drivers will have.

Actuarial Fairness By operating on a large scale, insurance companies can assure themselves that over a sufficiently large number of events, total premiums paid in will be equal to the total amount of money paid out. Let's return to our burglary example. A man knows that there is a 10-percent probability that his house will be burgled; if it is, he will suffer a \$10,000 loss. Prior to facing this risk, he calculates the expected loss to be \$1000 ($.10 \times \$10,000$). There is, however, substantial risk involved, because there is a 10-percent probability of a large loss. Now suppose that 100 people are similarly situated and that all of them buy burglary insurance from an insurance company. Because they all face a 10-percent probability of a \$10,000 loss, the insurance company might charge each of them a premium of \$1000. This \$1000 premium generates an insurance fund of \$100,000 from which losses can be paid. The insurance company can rely on the law of large numbers, which holds that the expected loss to the 100 individuals as a whole is likely to be very close to \$1000 each. The total payout, therefore, will be close to \$100,000, and the company need not worry about losing more than that.

When the insurance premium is equal to the expected payout, as in the example above, we say that the insurance is **actuarially fair**. Because they must cover administrative costs and make some profit, however, insurance companies typically charge premiums above expected losses. If there are a sufficient number of insurance companies to make the market competitive, these premiums will be close to actuarially fair levels. In some states, however, insurance premiums are regulated. Usually the objective is to protect consumers from "excessive" premiums. We will examine government regulation of markets in detail in Chapters 9 and 10 of this book.

actuarially fair Situation in which an insurance premium is equal to the expected payout.

EXAMPLE 5.3 The Value of Title Insurance When Buying a House

Suppose a family is buying its first house. They know that to close the sale, they'll need a deed that gives them clear "title." Without such a clear title, there is always a chance that the seller of the house is not its true owner. Of course, the seller could be engaging in fraud but is more likely to be unaware of the exact nature of his or her ownership rights. For example, the owner may have borrowed heavily, using the house as "collateral" for the loan. Or the property might carry with it a legal requirement that limits the use to which it may be put.

Suppose our family is willing to pay \$200,000 for the house but believes there is a one-in-twenty chance that careful research will reveal that the seller does not actually own the property. The property would then be worth nothing. If there were no insurance available, a risk-neutral family would bid at

TABLE 5.6 The Decision to Insure (\$)

INSURANCE	BURGLARY (PR = .1)	NO BURGLARY (PR = .9)	EXPECTED WEALTH	STANDARD DEVIATION
No	40,000	50,000	49,000	3,000
Yes	49,000	49,000	49,000	0

most \$190,000 for the property (.95[\$200,000] + .05[0]). However, a family that expects to tie up most of its assets in a house would probably be risk averse and, therefore, bid much less to buy the house—say, \$150,000.

In situations such as this, it is clearly in the interest of the buyer to be sure that there is no risk of a lack of full ownership. The buyer does this by purchasing “title insurance.” The title insurance company researches the history of the property, checks to see whether any legal liabilities are attached to it, and generally assures itself that there is no ownership problem. The insurance company then agrees to bear any remaining risk that might exist.

Because the title insurance company is a specialist in such insurance and can collect the relevant information relatively easily, the cost of title insurance is often less than the expected value of the loss involved. A fee of \$1,000 for title insurance is not unusual, and the expected loss can be substantially higher. It is also in the interest of sellers to provide title insurance, because all but the most risk-loving buyers will pay much more for the house when it is insured than when it is not. In fact, most states require sellers to provide title insurance before a sale can be completed. In addition, because mortgage lenders, too, are concerned about such risks, they usually require new buyers to have title insurance before they will issue a mortgage.

The Value of Information

People often make decisions based on limited information. If more information were available, one could make better predictions and reduce risk. Because information is a valuable commodity, people will pay for it. The **value of complete information** is the difference between the expected value of a choice when there is complete information and the expected value when information is incomplete.

value of complete information
Difference between the expected value of a choice when there is complete information and the expected value when information is incomplete.

To see how valuable information can be, suppose you are a store manager and must decide how many suits to order for the fall season. If you order 100 suits, your cost is \$180 per suit. If you order only 50 suits, your cost increases to \$200. You know that you will be selling suits for \$300 each, but you are not sure what total sales will be. All suits not sold can be returned, but for only half of what you paid for them. Without additional information, you will act on your belief that there is a .5 probability that 100 suits will be sold and a .5 probability that sales will be 50. Table 5.7 gives the profit that you would earn in each of these two cases.

Without additional information, you would choose to buy 100 suits if you were risk neutral, taking the chance that your profit might be either \$12,000 or \$1500. But if you were risk averse, you might buy 50 suits: In that case, you would know for sure that your profit would be \$5000.

	SALES OF 50	SALES OF 100	EXPECTED PROFIT
Buy 50 suits	5,000	5,000	5,000
Buy 100 suits	1,500	12,000	6,750

With complete information, you can place the correct order regardless of future sales. If sales were going to be 50 and you ordered 50 suits, your profits would be \$5000. If, on the other hand, sales were going to be 100 and you ordered 100 suits, your profits would be \$12,000. Because both outcomes are equally likely, your expected profit with complete information would be \$8500. The value of information is computed as

	Expected value with complete information:	\$8500
Less:	Expected value with uncertainty (buy 100 suits):	−\$6750
	Value of complete information	\$1750

Thus it is worth paying up to \$1750 to obtain an accurate prediction of sales. Even though forecasting is inevitably imperfect, it may be worth investing in a marketing study that provides a reasonable forecast of next year’s sales.

EXAMPLE 5.4 The Value of Information in the Dairy Industry

Historically, the U.S. dairy industry has allocated its advertising expenditures more or less uniformly throughout the year.⁹ But per capita consumption of milk has declined over the years—a situation that has stirred producers to look for new strategies to encourage milk consumption. One strategy would be to increase advertising expenditures and to continue advertising at a uniform rate throughout the year. A second strategy would be to invest in market research in order to obtain more information about the seasonal demand for milk; marketers could then reallocate expenditures so that advertising was most intense when the demand for milk was greatest.

Research into milk demand shows that sales follow a seasonable pattern, with demand greatest during the spring and lowest during the summer and early fall. The price elasticity of milk demand is negative but small and the income elasticity positive and large. Most important is the fact that milk advertising has the most effect on sales when consumers have the strongest preference for the product (March, April, and May) and least when preferences are weakest (August, September, and October).

In this case, the cost of obtaining seasonal information about milk demand is relatively low and the value of the information substantial. To estimate this value, we can compare the actual sales of milk during a typical year with sales levels that would have been reached had advertising expenditures been made in proportion to the strength of seasonal demand. In the latter case, 30 percent of the advertising budget would be allocated in the first quarter of the year and only 20 percent in the third quarter.

Making these calculations for the New York metropolitan area shows that the value of information—the value of the additional annual milk sales—was about \$4 million. This figure corresponds to a 9-percent increase in the profit to producers.

In §4.4, we define the price elasticity of demand as the percentage change in quantity demanded resulting from a 1-percent change in the price of a good.

⁹ This example is based on Henry Kinnucan and Olan D. Forker, “Seasonality in the Consumer Response to Milk Advertising with Implications for Milk Promotion Policy,” *American Journal of Agricultural Economics* 68 (1986): 562–71.

*5.4 The Demand for Risky Assets

Most people are risk averse. Given a choice, they prefer fixed monthly incomes to those which, though equally large on average, fluctuate randomly from month to month. Yet many of these same people will invest all or part of their savings in stocks, bonds, and other assets that carry some risk. Why do risk-averse people invest in the stock market and thereby risk losing part or all of their investments?¹⁰ How do people decide how much risk to bear when making investments and planning for the future? To answer these questions, we must examine the demand for risky assets.

Assets

asset Something that provides a flow of money or services to its owner.

An asset is *something that provides a flow of money or services to its owner*. A home, an apartment building, a savings account, or shares of General Motors stock are all assets. A home, for example, provides a flow of housing services to its owner, and if the owner did not wish to live there, could be rented out, thereby providing a monetary flow. Likewise, apartments in an apartment building can be rented out, providing a flow of rental income to the owner of the building. A savings account pays interest (usually every day or every month), which is usually reinvested in the account.

The monetary flow that one receives from asset ownership can take the form of an explicit payment, such as the rental income from an apartment building: Every month, the landlord receives rent checks from the tenants. Another form of explicit payment is the dividend on shares of common stock: Every three months the owner of a share of General Motors stock receives a quarterly dividend payment.

But sometimes the monetary flow from ownership of an asset is implicit: It takes the form of an increase or decrease in the price or value of the asset. An increase in the value of an asset is a *capital gain*, a decrease a *capital loss*. For example, as the population of a city grows, the value of an apartment building may increase. The owner of the building will then earn a capital gain beyond the rental income. The capital gain is *unrealized* until the building is sold because no money is actually received until then. There is, however, an implicit monetary flow because the building *could* be sold at any time. The monetary flow from owning General Motors stock is also partly implicit. The price of the stock changes from day to day, and each time it does, owners gain or lose.

Risky and Riskless Assets

risky asset Asset that provides an uncertain flow of money or services to its owner.

A **risky asset** provides a monetary flow that is at least in part random. In other words, the monetary flow is not known with certainty in advance. A share of General Motors stock is an obvious example of a risky asset: You cannot know whether the price of the stock will rise or fall over time, nor can you even be sure that the company will continue to pay the same (or any) dividend per share. Although people often associate risk with the stock market, most other assets are also risky.

¹⁰ Most Americans have at least some money invested in stocks or other risky assets, though often indirectly. For example, many people who hold full-time jobs have shares in pension funds underwritten in part by their own salary contributions and in part by employer contributions. Usually such funds are invested partly in the stock market.

An apartment building is one example. You cannot know how much land values will rise or fall, whether the building will be fully rented all the time, or even whether the tenants will pay their rents promptly. Corporate bonds are another example—the issuing corporation could go bankrupt and fail to pay bond owners their interest and principal. Even long-term U.S. government bonds that mature in 10 or 20 years are risky. Although it is highly unlikely that the federal government will go bankrupt, the rate of inflation could unexpectedly increase and make future interest payments and the eventual repayment of principal worth less in real terms, thereby reducing the value of the bonds.

In contrast, a **riskless (or risk-free) asset** pays a monetary flow that is known with certainty. Short-term U.S. government bonds—called Treasury bills—are riskless, or almost riskless. Because these bonds mature in a few months, there is very little risk from an unexpected increase in the rate of inflation. You can also be reasonably confident that the U.S. government will not default on the bond (i.e., refuse to pay back the holder when the bond comes due). Other examples of riskless or almost riskless assets include passbook savings accounts and short-term certificates of deposit.

riskless (or risk-free) asset Asset that provides a flow of money or services that is known with certainty.

Asset Returns

People buy and hold assets because of the monetary flows they provide. To compare assets with each other, it helps to think of this monetary flow relative to an asset's price or value. The **return** on an asset is *the total monetary flow it yields—including capital gains or losses—as a fraction of its price*. For example, a bond worth \$1000 today that pays out \$100 this year (and every year) has a return of 10 percent.¹¹ If an apartment building was worth \$10 million last year, increased in value to \$11 million this year, and also provided rental income (after expenses) of \$0.5 million, it would have yielded a return of 15 percent over the past year. If a share of General Motors stock was worth \$80 at the beginning of the year, fell to \$72 by the end of the year, and paid a dividend of \$4, it will have yielded a return of -5 percent (the dividend yield of 5 percent less the capital loss of 10 percent).

return Total monetary flow of an asset as a fraction of its price.

When people invest their savings in stocks, bonds, land, or other assets, they usually hope to earn a return that exceeds the rate of inflation, so that by delaying consumption, they could buy more in the future than they could by spending all their income now. Thus we often express the return on an asset in *real*—i.e., *inflation-adjusted*—terms. The **real return** on an asset is its simple (or nominal) return less the rate of inflation. For example, with an annual inflation rate of 5 percent, our bond, apartment building, and share of GM stock have yielded real returns of 5 percent, 10 percent, and -10 percent, respectively.

real return Simple (or nominal) return on an asset, less the rate of inflation.

Expected versus Actual Returns Because most assets are risky, an investor cannot know in advance what returns they will yield over the coming year. For example, our apartment building might have depreciated in value

¹¹ The price of a bond often changes during the course of a year. If the bond appreciates (or depreciates) in value during the year, its return will be greater (or less) than 10 percent. In addition, the definition of *return* given above should not be confused with the "internal rate of return," which is sometimes used to compare monetary flows occurring over some time. We discuss other return measures in Chapter 15, when we deal with present discounted values.

	REAL RATE OF RETURN (%)	RISK (STANDARD DEVIATION, %)
Common stocks (S&P 500)	9.5	20.2
Long-term corporate bonds	2.7	8.3
U.S. Treasury bills	0.6	3.2

expected return Return that an asset should earn on average.

actual return Return that an asset earns.

instead of appreciating, and the price of GM stock might have risen instead of falling. However, we can still compare assets by looking at their expected returns. The **expected return** on an asset is *the expected value of its return*, i.e., the return that it should earn on average. In some years, an asset's **actual return** may be much higher than its expected return and in some years much lower. Over a long period, however, the average return should be close to the expected return.

Different assets have different expected returns. Table 5.8, for example, shows that while the expected real return of a U.S. Treasury bill has been less than 1 percent, the expected real return on a group of representative stocks on the New York Stock Exchange has been more than 9 percent.¹² Why would anyone buy a Treasury bill when the expected return on stocks is so much higher? Because the demand for an asset depends not just on its expected return, but also on its *risk*. Although stocks have a higher expected return than Treasury bills, they also carry much more risk. One measure of risk, the standard deviation of the real annual return, is equal to 20.2 percent for common stocks, 8.3 percent for corporate bonds, and only 3.2 percent for U.S. Treasury bills.

The numbers in Table 5.8 suggest that the higher the expected return on an investment, the greater the risk involved. Assuming that one's investments are well diversified, this is indeed the case.¹³ As a result, the risk-averse investor must balance expected return against risk. We examine this trade-off in more detail in the next section.

The Trade-Off Between Risk and Return

Suppose a woman wants to invest her savings in two assets—Treasury bills, which are almost risk free, and a representative group of stocks.¹⁴ She must decide how much to invest in each asset. She might, for instance, invest only in

¹²For some stocks, the expected return is higher, and for some it is lower. Stocks of smaller companies (e.g., some of those traded on the NASDAQ) have higher expected rates of return—and higher return standard deviations.

¹³It is *nondiversifiable* risk that matters. An individual stock may be very risky but still have a low expected return because most of the risk could be diversified away by holding a large number of such stocks. *Nondiversifiable risk*, which arises from the fact that individual stock prices are correlated with the overall stock market, is the risk that remains even if one holds a diversified portfolio of stocks. We discuss this point in detail in the context of the *Capital Asset Pricing Model* in Chapter 15.

¹⁴The easiest way to invest in a representative group of stocks is to buy shares in a *mutual fund*. Because a mutual fund invests in many stocks, one effectively buys a portfolio.

Treasury bills, only in stocks, or in some combination of the two. As we will see, this problem is analogous to the consumer's problem of allocating a budget between purchases of food and clothing.

Let's denote the risk-free return on the Treasury bill by R_f . Because the return is risk free, the expected and actual returns are the same. In addition, let the *expected* return from investing in the stock market be R_m and the actual return be r_m . The actual return is risky. At the time of the investment decision, we know the set of possible outcomes and the likelihood of each, but we do not know what particular outcome will occur. The risky asset will have a higher expected return than the risk-free asset ($R_m > R_f$). Otherwise, risk-averse investors would buy only Treasury bills and no stocks would be sold.

The Investment Portfolio To determine how much money the investor should put in each asset, let's set b equal to the fraction of her savings placed in the stock market and $(1 - b)$ the fraction used to purchase Treasury bills. The expected return on her total portfolio, R_p , is a weighted average of the expected return on the two assets:¹⁵

$$R_p = bR_m + (1 - b)R_f \quad (5.1)$$

Suppose, for example, that Treasury bills pay 4 percent ($R_f = .04$), the stock market's expected return is 12 percent ($R_m = .12$), and $b = 1/2$. Then $R_p = 8$ percent. How risky is this portfolio? One measure of its riskiness is the standard deviation of its return. We will denote the *standard deviation* of the risky stock market investment by σ_m . With some algebra, we can show that the *standard deviation of the portfolio*, σ_p (with one risky and one risk-free asset) is the fraction of the portfolio invested in the risky asset times the standard deviation of that asset:¹⁶

$$\sigma_p = b\sigma_m \quad (5.2)$$

The Investor's Choice Problem

We have still not determined how the investor should choose this fraction b . To do so, we must first show that she faces a risk-return trade-off analogous to a consumer's budget line. To identify this trade-off, note that equation (5.1) for the expected return on the portfolio can be rewritten as

$$R_p = R_f + b(R_m - R_f)$$

¹⁵The expected value of the sum of two variables is the sum of the expected values. Therefore

$$R_p = E[br_m] + E[(1 - b)R_f] = bE[r_m] + (1 - b)R_f = bR_m + (1 - b)R_f$$

¹⁶To see why, we observe from footnote 4 that we can write the variance of the portfolio return as

$$\sigma_p^2 = E[br_m + (1 - b)R_f - R_p]^2$$

Substituting equation (5.1) for the expected return on the portfolio, R_p , we have

$$\sigma_p^2 = E[br_m + (1 - b)R_f - bR_m - (1 - b)R_f]^2 = E[b(r_m - R_m)]^2 = b^2\sigma_m^2$$

Because the standard deviation of a random variable is the square root of its variance, $\sigma_p = b\sigma_m$.

In §3.2, we explain how a budget line is determined from an individual's income and the prices of the available goods.

Now, from equation (5.2) we see that $b = \sigma_p / \sigma_m$, so that

$$R_p = R_f + \frac{(R_m - R_f)}{\sigma_m} \sigma_p \quad (5.3)$$

Risk and the Budget Line This equation is a *budget line* because it describes the trade-off between risk (σ_p) and expected return (R_p). Note that it is the equation for a straight line: Because R_m , R_f , and σ_m are constants, the slope $(R_m - R_f) / \sigma_m$ is a constant, as is the intercept R_f . The equation says that the expected return on the portfolio R_p increases as the standard deviation of that return σ_p increases. We call the slope of this budget line, $(R_m - R_f) / \sigma_m$, the **price of risk** because it tells us how much extra risk an investor must incur to enjoy a higher expected return.

price of risk Extra risk that an investor must incur to enjoy a higher expected return.

The budget line is drawn in Figure 5.6. If our investor wants no risk, she can invest all her funds in Treasury bills ($b = 0$) and earn an expected return R_f . To receive a higher expected return, she must incur some risk. For example, she could invest all her funds in stocks ($b = 1$), earning an expected return R_m but

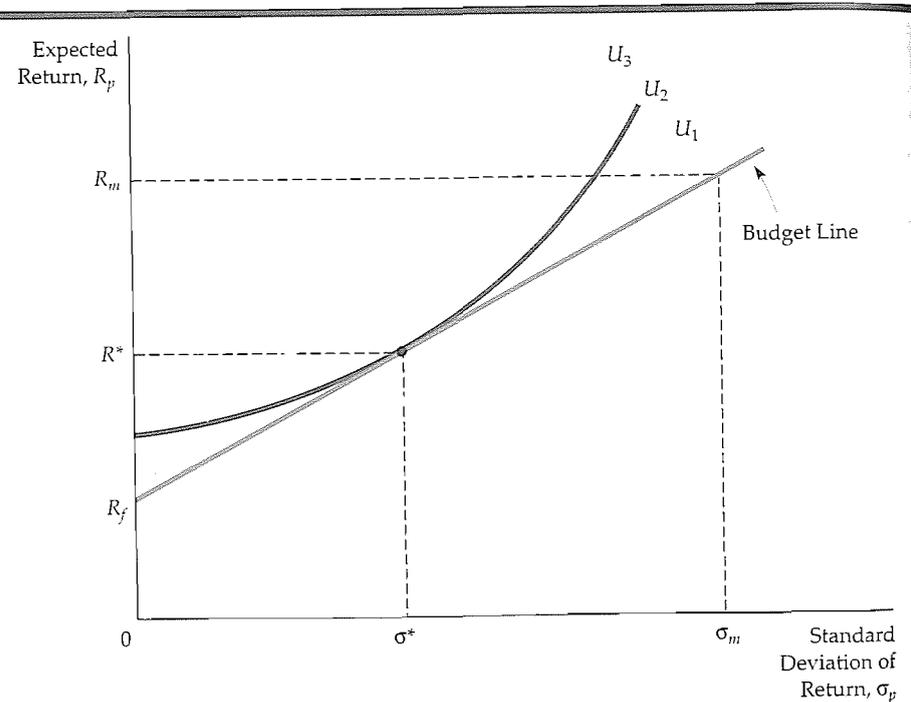


FIGURE 5.6 Choosing Between Risk and Return

An investor is dividing her funds between two assets—Treasury bills, which are risk free, and stocks. The budget line describes the trade-off between the expected return and its riskiness, as measured by its standard deviation. The slope of the budget line is $(R_m - R_f) / \sigma_m$, which is the price of risk. Three indifference curves are drawn, each showing combinations of risk and return that leave an investor equally satisfied. The curves are upward-sloping because a risk-averse investor will require a higher expected return if she is to bear a greater amount of risk. The utility-maximizing investment portfolio is at the point where indifference curve U_2 is tangent to the budget line.

incurring a standard deviation σ_m . Or she might invest some fraction of her funds in each type of asset, earning an expected return somewhere between R_f and R_m and facing a standard deviation less than σ_m but greater than zero.

Risk and Indifference Curves. Figure 5.6 also shows the solution to the investor's problem. Three indifference curves are drawn in the figure. Each curve describes combinations of risk and return that leave the investor equally satisfied. The curves are upward-sloping because risk is undesirable. Thus with a greater amount of risk, it takes a greater expected return to make the investor equally well-off. The curve U_3 yields the greatest amount of satisfaction and U_1 the least amount: For a given amount of risk, the investor earns a higher expected return on U_3 than on U_2 , and a higher expected return on U_2 than on U_1 .

Of the three indifference curves, the investor would prefer to be on U_3 . This position, however, is not feasible, because U_3 does not touch the budget line. Curve U_1 is feasible, but the investor can do better. Like the consumer choosing quantities of food and clothing, our investor does best by choosing a combination of risk and return at the point where an indifference curve (in this case U_2) is tangent to the budget line. At that point, the investor's return has an expected value R^* and a standard deviation σ^* .

Naturally, people differ in their attitudes toward risk. This fact is illustrated in Figure 5.7, which shows how two different investors choose their portfolios.

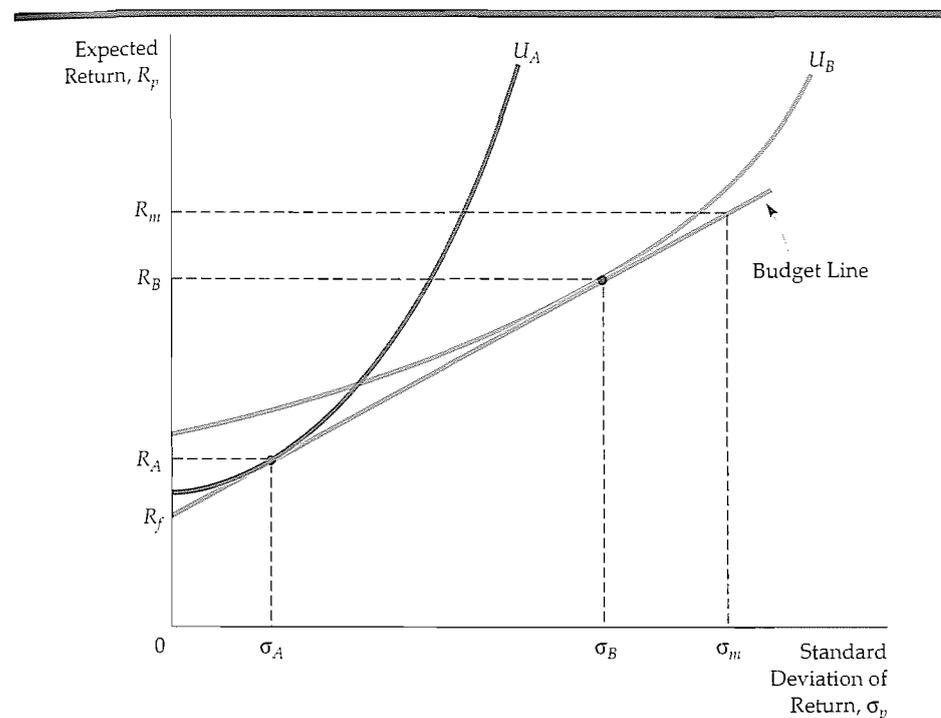


FIGURE 5.7 The Choices of Two Different Investors

Investor A is highly risk averse. Because his portfolio will consist mostly of the risk-free asset, his expected return R_A will be only slightly greater than the risk-free return. His risk σ_A , however, will be small. Investor B is less risk averse. She will invest a large fraction of her funds in stocks. Although the expected return on her portfolio R_B will be larger, the return will also be riskier.

Investor A is quite risk averse. Because his indifference curve U_A is tangent to the budget line at a point of low risk, he will invest almost all his funds in Treasury bills and earn an expected return R_A just slightly larger than the risk-free return R_f . Investor B is less risk averse. She will invest most of her funds in stocks, and while the return on her portfolio will have a higher expected value R_B , it will also have a higher standard deviation σ_B .

If Investor B has a sufficiently low level of risk aversion, she might buy stocks on margin: that is, she would borrow money from a brokerage firm in order to invest more than she actually owns in the stock market. In effect, a person who buys stocks on margin holds a portfolio with more than 100 percent of the portfolio's value invested in stocks. This situation is illustrated in Figure 5.8, which shows indifference curves for two investors. Investor A , who is relatively risk-averse, invests about half of his funds in stocks. Investor B , however, has an indifference curve that is relatively flat and tangent with the budget line at a point where the expected return on the portfolio exceeds the expected return on the stock market. In order to hold this portfolio, the investor must borrow money because she wants to invest more than 100 percent of her wealth in the stock market. Buying stocks on margin in this way is a form of *leverage*: the investor increases her expected return above that for the overall stock market, but at the cost of increased risk.

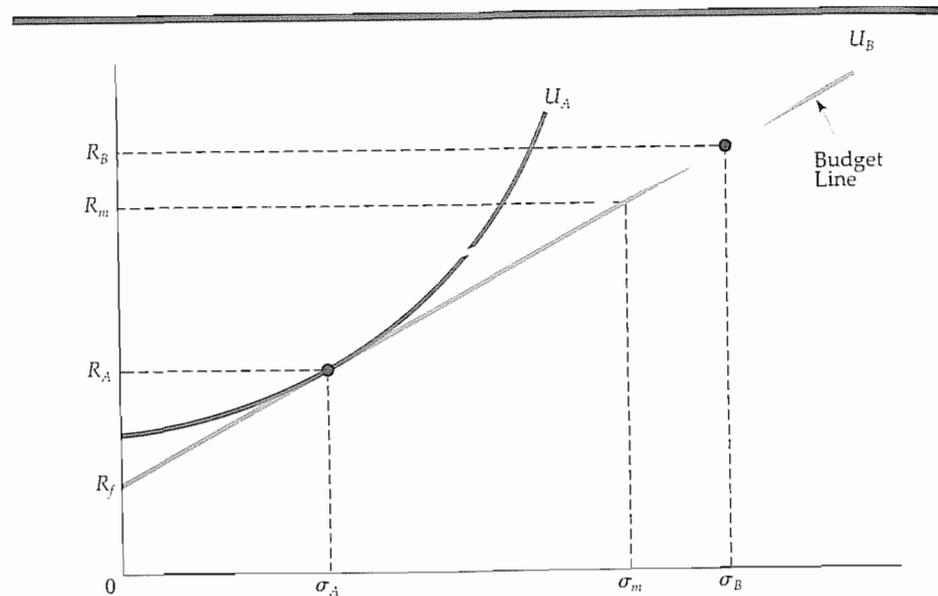


FIGURE 5.8 Buying Stocks on Margin

Because Investor A is risk averse, his portfolio contains a mixture of stocks and risk-free Treasury bills. Investor B , however, has a very low degree of risk aversion. Her indifference curve, U_B , is tangent to the budget line at a point where the expected return and standard deviation for her portfolio exceed those for the stock market overall. This implies that she would like to invest more than 100 percent of her wealth in the stock market. She does so by buying stocks on margin—i.e., by borrowing from a brokerage firm to help finance the investment.

In Chapters 3 and 4, we simplified the problem of consumer choice by assuming that the consumer had only two goods from which to choose—food and clothing. In the same spirit, we have simplified the investor's choice by limiting it to Treasury bills and stocks. The basic principles, however, would be the same if we had more assets (e.g., corporate bonds, land, and different types of stocks). Every investor faces a trade-off between risk and return.¹⁷ The degree of extra risk that each is willing to bear in order to earn a higher expected return depends on how risk averse he or she is. Less risk-averse investors tend to include a larger fraction of risky assets in their portfolios.

EXAMPLE 5.5 Investing in the Stock Market

During the 1990s, we witnessed a shift in the investing behavior of Americans. First, many Americans started investing in the stock market for the first time. In 1989, about 32 percent of families in the United States had part of their wealth invested in the stock market, either directly (by owning individual stocks) or indirectly (through mutual funds or pension plans invested in stocks). By 1995, that fraction had risen to above 41 percent. In addition, the share of wealth invested in stocks increased from about 26 percent to about 40 percent during this period.¹⁸

Much of this shift is attributable to younger investors. For those under the age of 35, participation in the stock market increased from about 23 percent in 1989 to about 39 percent in 1995. For those older than 35, participation also increased, though by much less.

Why have more people, and especially younger people, started investing in the stock market? One reason is the advent of on-line trading over the Internet, which has made investing much easier. Another reason may be the considerable increase in stock prices that occurred during the late 1990s. These increases may have convinced some investors that prices could only continue to rise in the future. As one analyst has put it, "The market's relentless seven-year climb, the popularity of mutual funds, the shift by employers to self-directed retirement plans, and the avalanche of do-it-yourself investment publications all have combined to create a nation of financial know-it-alls."¹⁹

The run-up in the stock market during the 1990s has indeed surprised many people. Although the American economy has been very strong over this period, by 1999 prices reached almost unprecedented levels relative to earnings and dividends. Figure 5.9 shows the dividend yield and price/earnings ratio for the S&P 500 (an index of the stocks of 500 large corporations) over the period 1980–1999. Observe that the dividend yield (the annual dividend divided by the stock price) fell from about 5 percent in 1980 to about 1.5 percent in 1999. The

¹⁷ As mentioned earlier, what matters is nondiversifiable risk, because investors can eliminate diversifiable risk by holding many different stocks (e.g., via mutual funds). We discuss diversifiable versus nondiversifiable risk in Chapter 15.

¹⁸ Data are from the *Federal Reserve Bulletin*, January 1997.

¹⁹ "We're All Bulls Here: Strong Market Makes Everybody an Expert," *Wall Street Journal*, September 12, 1997.

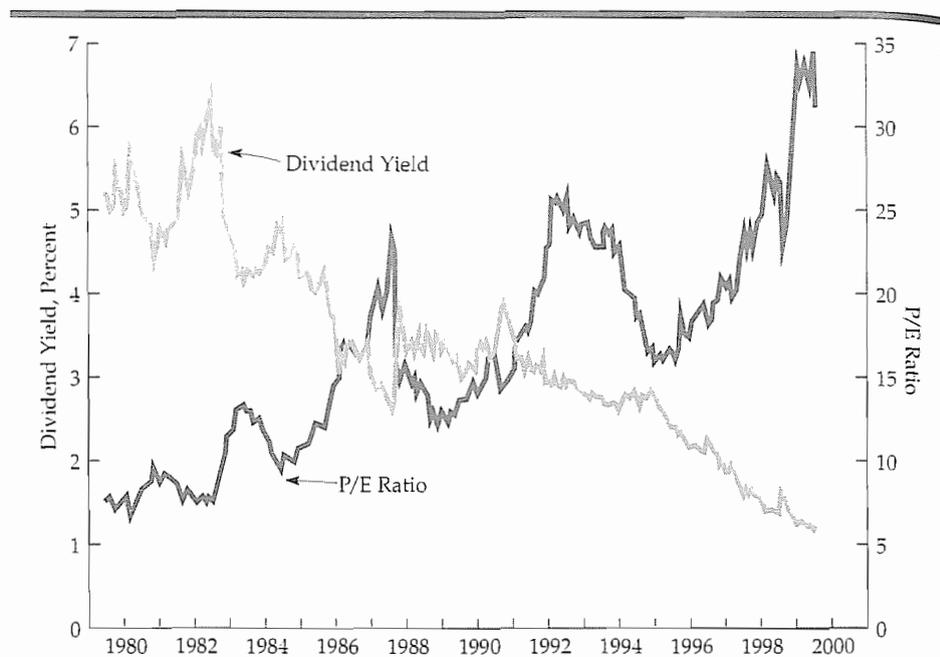


FIGURE 5.9 Dividend Yield and P/E Ratio for S&P 500

The dividend yield (the annual dividend divided by the stock price) fell dramatically from 1980 through 1999, while the price/earnings ratio (the stock price divided by the annual earnings-per-share) rose on average for the S&P 500.

price/earnings ratio (the share price divided by the annual earnings-per-share) increased from about 8 to nearly 35. These ratios would only be justified if one believed that corporate profits will continue to grow rapidly over the coming decade. This situation suggests that in the late 1990s, many investors had a low degree of risk aversion and/or were quite optimistic about the economy.

SUMMARY

1. Consumers and managers frequently make decisions in which there is uncertainty about the future. This uncertainty is characterized by the term *risk*, which applies when each of the possible outcomes and its probability of occurrence is known.
2. Consumers and investors are concerned about the expected value and the variability of uncertain outcomes. The expected value is a measure of the central tendency of the value of the risky outcomes. The variability is frequently measured by the standard deviation of outcomes, which is the square root of the average of the squares of the deviations of each possible outcome from its expected value.
3. Facing uncertain choices, consumers maximize their expected utility—an average of the utility associated with each outcome—with the associated probabilities serving as weights.
4. A person who would prefer a certain return of a given amount to a risky investment whose expected return is the same amount is risk averse. The maximum amount of money that a risk-averse person would pay to avoid taking a risk is called the *risk premium*. A person who is indifferent between a risky investment and the certain receipt of the expected return on that investment is risk neutral. A risk-loving consumer would prefer a risky investment with a given

expected return to the certain receipt of that expected return.

5. Risk can be reduced by (a) diversification, (b) insurance, and (c) obtaining additional information.
6. The *law of large numbers* enables insurance companies to provide insurance for which the premium paid

equals the expected value of the loss being insured against. We call such insurance *actuarially fair*.

7. Consumer theory can be applied to decisions to invest in risky assets. The budget line reflects the price of risk, and consumers' indifference curves reflect their attitudes toward risk.

QUESTIONS FOR REVIEW

1. What does it mean to say that a person is *risk averse*? Why are some people likely to be risk averse while others are risk lovers?
2. Why is the variance a better measure of variability than the range?
3. What does it mean for consumers to maximize expected utility? Can you think of a case in which a person might *not* maximize expected utility?
4. Why do people often want to insure fully against uncertain situations even when the premium paid exceeds the expected value of the loss being insured against?
5. Why is an insurance company likely to behave as if it were risk neutral even if its managers are risk-averse individuals?
6. When is it worth paying to obtain more information to reduce uncertainty?
7. How does the diversification of an investor's portfolio avoid risk?
8. Why do some investors put a large portion of their portfolios into risky assets while others invest largely in risk-free alternatives? (*Hint*: Do the two investors receive exactly the same return on average? If so, why?)

EXERCISES

1. Consider a lottery with three possible outcomes:
 - \$100 will be received with probability .1
 - \$50 will be received with probability .2
 - \$10 will be received with probability .7
 - a. What is the expected value of the lottery?
 - b. What is the variance of the outcomes?
 - c. What would a risk-neutral person pay to play the lottery?
2. Suppose you have invested in a new computer company whose profitability depends on two factors: (1) whether the U.S. Congress passes a tariff raising the cost of Japanese computers and (2) whether the U.S. economy grows slowly or quickly. What are the four mutually exclusive states of the world that you should be concerned about?
3. Richard is deciding whether to buy a state lottery ticket. Each ticket costs \$1, and the probability of winning payoffs is given as follows:

PROBABILITY	RETURN
.5	\$0.00
.25	\$1.00
.2	\$2.00
.05	\$7.50

 - a. What is the expected value of Richard's payoff if he buys a lottery ticket? What is the variance?
 - b. Richard's nickname is "No-Risk Rick" because he is an extremely risk-averse individual. Would he buy the ticket?
 - c. Suppose Richard was offered insurance against losing any money. If he buys 1,000 lottery tickets, how much would he be willing to pay to insure his gamble?
 - d. In the long run, given the price of the lottery ticket and the probability/return table, what do you think the state would do about the lottery?
4. Suppose an investor is concerned about a business choice in which there are three prospects—the probability and returns are given below:

PROBABILITY	RETURN
.2	\$100
.4	50
.4	-25

What is the expected value of the uncertain investment? What is the variance?

5. You are an insurance agent who must write a policy for a new client named Sam. His company, Society for Creative Alternatives to Mayonnaise (SCAM), is working on a low-fat, low-cholesterol mayonnaise substitute for the sandwich-condiment industry. The sandwich industry will pay top dollar to the first inventor to patent such a mayonnaise substitute. Sam's SCAM seems like a very risky proposition to you. You have calculated his possible returns table as follows:

PROBABILITY	RETURN	
.999	-\$1,000,000	(he fails)
.001	\$1,000,000,000	(he succeeds and sells his formula)

- a. What is the expected return of Sam's project? What is the variance?
- b. What is the most that Sam is willing to pay for insurance? Assume Sam is risk neutral.
- c. Suppose you found out that the Japanese are on the verge of introducing their own mayonnaise substitute next month. Sam does not know this and has just turned down your final offer of \$1000 for the insurance. Assume that Sam tells you SCAM is only six months away from perfecting its mayonnaise substitute *and* that you know what you know about the Japanese. Would you raise or lower your policy premium on any subsequent proposal to Sam? Based on his information, would Sam accept?
6. Suppose that Natasha's utility function is given by $u(I) = I^{0.5}$, where I represents annual income in thousands of dollars.
- a. Is Natasha risk loving, risk neutral, or risk averse? Explain.
- b. Suppose that Natasha is currently earning an income of \$10,000 ($I = 10$) and can earn that income next year with certainty. She is offered a chance to take a new job that offers a .5 probability of earning \$16,000 and a .5 probability of earning \$5000. Should she take the new job?
- c. In (b), would Natasha be willing to buy insurance to protect against the variable income associated with the new job? If so, how much would she be willing to pay for that insurance? (*Hint*: What is the risk premium?)
7. Draw a utility function over income $u(I)$ that describes a man who is a risk lover when his income is low but a risk averter when his income is high. Can you explain why such a utility function might reasonably describe a person's preferences?
8. A city is considering how much to spend to monitor its parking meters. The following information is available to the city manager:
- Hiring each meter monitor costs \$10,000 per year.
 - With one monitoring person hired, the probability of a driver getting a ticket each time he or she parks illegally is equal to .25.
 - With two monitors hired, the probability of getting a ticket is .5; with three monitors, the probability is .75; and with four it's equal to 1.
 - With two metering persons hired, the current fine for overtime parking is \$20.
- a. Assume first that all drivers are risk neutral. What parking fine would you levy and how many meter monitors would you hire (1, 2, 3, or 4) to achieve the current level of deterrence against illegal parking at the minimum cost?
- b. Now assume that drivers are highly risk averse. How would your answer to (a) change?
- c. (For discussion) What if drivers could insure themselves against the risk of parking fines? Would it make good public policy to allow such insurance to be available?
9. A moderately risk-averse investor has 50 percent of her portfolio invested in stocks and 50 percent invested in risk-free Treasury bills. Show how each of the following events will affect the investor's budget line and the proportion of stocks in her portfolio:
- a. The standard deviation of the return on the stock market increases, but the expected return on the stock market remains the same.
- b. The expected return on the stock market increases, but the standard deviation of the stock market remains the same.
- c. The return on risk-free Treasury bills increases.

CHAPTER 6

Production

In the last three chapters, we focused on the *demand side* of the market—the preferences and behavior of consumers. Now we turn to the *supply side* and examine the behavior of producers. We will see how firms can produce efficiently and how their costs of production change with changes in both input prices and the level of output. We will also see that there are strong similarities between the optimizing decisions made by firms and those made by consumers. In other words, understanding consumer behavior will help us understand producer behavior.

In this chapter and the next we discuss the **theory of the firm**, which describes how a firm makes cost-minimizing production decisions, and how the firm's resulting cost varies with its output. Our knowledge of production and cost will help us understand the characteristics of market supply. It will also prove useful for dealing with problems that arise regularly in business. To see this, just consider some of the problems often faced by a company like General Motors. How much assembly-line machinery and how much labor should it use in its new automobile plants? If it wants to increase production, should it hire more workers, construct new plants, or both? Does it make more sense for one automobile plant to produce different models, or should each model be manufactured in a separate plant? What should GM expect its costs to be during the coming year? How are these costs likely to change over time and be affected by the level of production? These questions apply not only to business firms but also to other producers of goods and services, such as governments and nonprofit agencies.

In this chapter we study the firm's *production technology*: the physical relationship that describes how inputs (such as labor and capital) are transformed into outputs (such as cars and televisions). We do this in several steps. First, we show how the production technology can be represented in the form of a *production function*—a compact description of how inputs are turned into output. Then, we use the production function to show how the firm's output changes when first one and then all inputs are varied. We will be particularly concerned with the *scale* of the firm's operation. For example, are there technological advantages that make the firm more productive as its scale increases?

Chapter Outline

- 6.1 The Technology of Production 178
- 6.2 Isoquants 179
- 6.3 Production with One Variable Input (Labor) 181
- 6.4 Production with Two Variable Inputs 191
- 6.5 Returns to Scale 197

List of Examples

- 6.1 Malthus and the Food Crisis 187
- 6.2 Labor Productivity and the Standard of Living 189
- 6.3 A Production Function for Wheat 196
- 6.4 Returns to Scale in the Carpet Industry 199

6.1 The Technology of Production

theory of the firm Explanation of how a firm makes cost-minimizing production decisions and how its cost varies with its output.

factors of production Inputs into the production process (e.g., labor, capital, and materials).

In the production process, firms turn *inputs* into *outputs* (or products). Inputs, which are also called **factors of production**, include anything that the firm must use as part of the production process. For example, for a bakery, inputs include the labor of its workers; raw materials, such as flour and sugar; and the capital invested in its ovens, mixers, and other equipment to produce such outputs as bread, cakes, and pastries.

We can divide inputs into the broad categories of *labor*, *materials*, and *capital*, each of which might include more narrow subdivisions. Labor inputs include skilled workers (carpenters, engineers) and unskilled workers (agricultural workers), as well as the entrepreneurial efforts of the firm's managers. Materials include steel, plastics, electricity, water, and any other goods that the firm buys and transforms into final products. Capital includes buildings, machinery and other equipment, and inventories.

The Production Function

The relationship between the inputs to the production process and the resulting output is described by a production function. A **production function** indicates the highest output Q that a firm can produce for every specified combination of inputs. For simplicity, we will assume that there are two inputs, labor L and capital K . We can then write the production function as

$$Q = F(K, L) \quad (6.1)$$

This equation relates the quantity of output to the quantities of the two inputs, capital and labor. For example, the production function might describe the number of personal computers that can be produced each year with a 10,000-square-foot plant and a specific amount of assembly-line labor. Or it might describe the crop that a farmer can obtain using a specific amount of machinery and workers.

It is important to keep in mind that inputs and outputs are *flows*. For example, a personal computer manufacturer uses a certain amount of labor *each year* to produce some number of computers over that year. Although the firm might own its plant and machinery, we can think of the firm as paying a cost for the use of that plant and machinery over the year. To simplify things, we will frequently ignore the reference to time and refer only to amounts of labor, capital, and output. Unless otherwise indicated, however, we mean the amount of labor and capital used each year and the amount of output produced each year.

The production function allows inputs to be combined in varying proportions, so that output can be produced in many ways. For the production function in equation (6.1), this could mean using more capital and less labor, or vice versa. For example, wine can be produced in a labor-intensive way using many workers, or in a capital-intensive way using machines and only a few workers.

Note that equation (6.1) applies to a *given technology*—that is, a given state of knowledge about the various methods that might be used to transform inputs into outputs. As the technology becomes more advanced and the production function changes, a firm can obtain more output for a given set of inputs. For example, a new, faster assembly-line may allow a hardware manufacturer to produce more high-speed computers in a given period of time.

Production functions describe what is *technically feasible* when the firm operates *efficiently*—that is, when the firm uses each combination of inputs as effectively as possible. The presumption that production is always technically efficient need not always hold, but it is reasonable to expect that profit-seeking firms will not waste resources.

6.2 Isoquants

Let's begin by examining the production technology of a firm that uses two inputs and can vary both of them. Suppose that the inputs are labor and capital and that they are used to produce food. Table 6.1 tabulates the output achievable for various combinations of inputs.

Labor inputs are listed across the top row, capital inputs down the column on the left. Each entry in the table is the maximum (technically efficient) output that can be produced each year with each combination of labor and capital used over that year. For example, 4 units of labor per year and 2 units of capital per year yield 85 units of food per year. Reading along each row, we see that output increases as labor inputs are increased, while capital inputs remain fixed. Reading down each column, we see that output also increases as capital inputs are increased, while labor inputs remain fixed.

The information contained in Table 6.1 can also be represented graphically using isoquants. An **isoquant** is a curve that shows all the possible combinations of inputs that yield the same output. Figure 6.1 shows three isoquants. (Each axis in the figure measures the quantity of inputs.) These isoquants are based on the data in Table 6.1, but have been drawn as smooth curves to allow for the use of fractional amounts of inputs.

For example, isoquant Q_1 shows all combinations of labor and capital per year that together yield 55 units of output per year. Two of these points, A and D , correspond to Table 6.1. At A , 1 unit of labor and 3 units of capital yield 55 units of output; at D , the same output is produced from 3 units of labor and 1 unit of capital. Isoquant Q_2 shows all combinations of inputs that yield 75 units of output and corresponds to the four combinations of labor and capital circled in the table (e.g., at B , where 2 units of labor and 3 units of capital are combined). Isoquant Q_2 lies above and to the right of Q_1 because obtaining a higher level of output requires more labor and capital. Finally, isoquant Q_3 shows labor-capital combinations that yield 90 units of output. Point C involves 3 units of labor and 3 units of capital, whereas Point E involves 2 units of labor and 5 units of capital.

isoquant Curve showing all possible combinations of inputs that yield the same output.

TABLE 6.1 Production with Two Variable Inputs

CAPITAL INPUT	LABOR INPUT				
	1	2	3	4	5
1	20	40	55	65	(75)
2	40	60	(75)	85	90
3	55	(75)	90	100	105
4	65	85	100	110	115
5	(75)	90	105	115	120

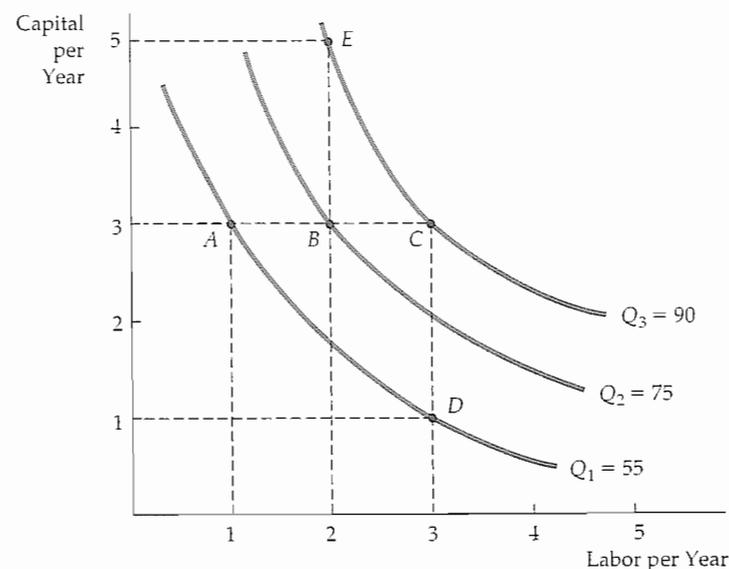


FIGURE 6.1 Production with Two Variable Inputs

Production isoquants show the various combinations of inputs necessary for the firm to produce a given output. A set of isoquants, or *isoquant map*, describes the firm's production function. Output increases as we move from isoquant Q_1 (at which 55 units per year are produced at points such as A and D), to isoquant Q_2 (75 units per year at points such as B) and to isoquant Q_3 (90 units per year at points such as C and E).

isoquant map Graph combining several isoquants, used to describe a production function.

Isoquant Maps When several isoquants are combined together in a single graph, as in Figure 6.1, we call the graph an **isoquant map**. An isoquant map is another way of describing a production function, just as an indifference map is a way of describing a utility function. Each isoquant corresponds to a different level of output, and the level of output increases as we move up and to the right in the figure.

Input Flexibility

Isoquants show the flexibility that firms have when making production decisions: They can usually obtain a particular output by substituting one input for another. It is important for the managers to understand the nature of this flexibility. For example, fast-food restaurants have recently faced shortages of young, low-wage employees. Companies have responded by automating—adding self-service salad bars and introducing more sophisticated cooking equipment. They have also recruited older people to fill positions. As we will see in Chapters 7 and 8, by taking this flexibility in the production process into account, managers can choose input combinations that minimize cost and maximize profit.

The Short Run versus the Long Run

The isoquants in Figure 6.1 show how capital and labor can be substituted for each other to produce the same amount of output. In practice, however, this substitution can take time. A new factory must be planned and built, and machinery and other capital equipment must be ordered and delivered. These activities can

easily take a year or more to complete. As a result, if we are looking at production decisions over a short period of time, such as a month or two, the firm is unlikely to be able to substitute very much capital for labor.

Because firms must consider whether or not inputs can be varied, and if they can, over what period of time, it is important to distinguish between the short and long run when analyzing production. The **short run** refers to a period of time in which one or more factors of production cannot be changed. In other words, in the short run there is at least one factor that cannot be varied; such a factor is called a **fixed input**. The **long run** is the amount of time needed to make all inputs variable.

As you might expect, the kinds of decisions that firms can make are very different in the short run than in the long run. In the short run, firms vary the intensity with which they utilize a given plant and machinery; in the long run, they vary the size of the plant. All fixed inputs in the short run represent the outcomes of previous long-run decisions based on estimates of what a firm could profitably produce and sell.

There is no specific time period, such as one year, that separates the short run from the long run. Rather, one must distinguish them on a case-by-case basis. For example, the long run can be as brief as a day or two for a child's lemonade stand, or as long as five or ten years for a petrochemical producer or an automobile manufacturer.

short run Period of time in which quantities of one or more production factors cannot be changed.

fixed input Production factor that cannot be varied.

long run Amount of time needed to make all production inputs variable.

6.3 Production with One Variable Input (Labor)

When deciding how much of a particular input to buy, a firm has to compare the benefit that will result with the cost. Sometimes it is useful to look at the benefit and the cost on an *incremental* basis by focusing on the additional output that results from an incremental addition to an input. In other situations it is useful to make the comparison on an *average* basis by considering the result of substantially increasing an input. We will look at these benefits and costs in both ways.

Let's begin by considering the case in which capital is fixed but labor is variable. (Because one of the factors is fixed, this is a short-run analysis.) In this case, the only way the firm can produce more output is by increasing its labor input. Imagine, for example, that you are managing a clothing factory. Although you have a fixed amount of equipment, you can hire more or less labor to sew and to run the machines. You must decide how much labor to hire and how much clothing to produce. To make the decision, you will need to know how the amount of output Q increases (if at all) as the input of labor L increases.

Table 6.2 gives this information. The first three columns show the amount of output that can be produced in one month with different amounts of labor, and capital fixed at 10 units. The first column shows the amount of labor, the second the fixed amount of capital, and the third total output. When labor input is zero, output is also zero. Output then increases as labor is increased up to an input of 8 units. Beyond that point, total output declines: Although initially each unit of labor can take greater and greater advantage of the existing machinery and plant, after a certain point, additional labor is no longer useful and indeed can be counterproductive. Five people can run an assembly line better than two, but ten people may get in each other's way.

TABLE 6.2 Production with One Variable Input				
AMOUNT OF LABOR (L)	AMOUNT OF CAPITAL (K)	TOTAL OUTPUT (Q)	AVERAGE PRODUCT (Q/L)	MARGINAL PRODUCT ($\Delta Q/\Delta L$)
0	10	0	—	—
1	10	10	10	10
2	10	30	15	20
3	10	60	20	30
4	10	80	20	20
5	10	95	19	15
6	10	108	18	13
7	10	112	16	4
8	10	112	14	0
9	10	108	12	-4
10	10	100	10	-8

Average and Marginal Products

The contribution that labor makes to the production process can be described on both an *average* and a *marginal* (i.e., incremental) basis. The fourth column in Table 6.2 shows the **average product of labor** (AP_L), which is the output per unit of labor input. The average product is calculated by dividing the total output Q by the total input of labor L . The average product of labor measures the productivity of the firm's workforce in terms of how much output each worker produces on average. In our example the average product increases initially but falls when the labor input becomes greater than four.

average product Output per unit of a particular input.

marginal product Additional output produced as an input is increased by one unit.

The fifth column of Table 6.2 shows the **marginal product of labor** (MP_L). This is the *additional* output produced as the labor input is increased by 1 unit. For example, with capital fixed at 10 units, when the labor input increases from 2 to 3, total output increases from 30 to 60, creating an additional output of 30 (i.e., $60 - 30$) units. The marginal product of labor can be written as $\Delta Q/\Delta L$, in other words the change in output ΔQ resulting from a 1-unit increase in labor input ΔL .

Remember that the marginal product of labor depends on the amount of capital used. If the capital input increased from 10 to 20, the marginal product of labor would most likely increase. Why? Because additional workers are likely to be more productive if they have more capital to use. Like the average product, the marginal product first increases then falls, in this case after the third unit of labor.

To summarize:

<p>Average product of labor = Output/labor input = Q/L</p> <p>Marginal product of labor = Change in output/change in labor input = $\Delta Q/\Delta L$</p>
--

The Slopes of the Product Curve

Figure 6.2 plots the information contained in Table 6.2. (We have connected all the points in the figure with solid lines.) Figure 6.2(a) shows that as labor is increased output increases until it reaches the maximum output of 112;

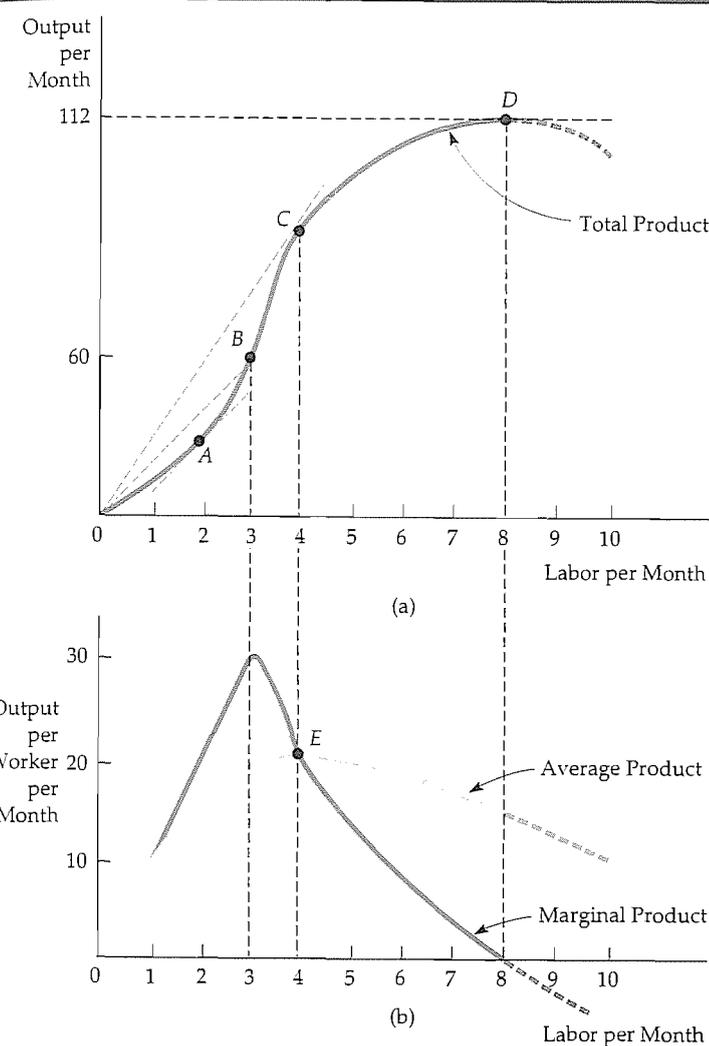


FIGURE 6.2 Production with One Variable Input

The total product curve in (a) shows the output produced for different amounts of labor input. The average and marginal products in (b) can be obtained (using the data in Table 6.2) from the total product curve. At point A, the marginal product is 20 because the tangent to the total product curve has a slope of 20. At point B in (a) the average product of labor is 20, which is the slope of the line from the origin to B. The average product of labor at point C in (a) is given by the slope of the line OC. To the left of point E in (b) the marginal product is above the average product and the average is increasing; to the right of E the marginal product is below the average product and the average is decreasing. As a result, E represents the point at which the average and marginal products are equal, when the average product reaches its maximum.

thereafter it falls. The portion of the total output curve that is declining is drawn with a dashed line to denote that producing with more than eight workers is not economically rational; it can never be profitable to use additional amounts of a costly input to produce *less* output.

Figure 6.2(b) shows the average and marginal product curves. (The units on the vertical axis have changed from output per month to output per worker per month.) Note that the marginal product is positive as long as output is increasing, but becomes negative when output is decreasing.

It is no coincidence that the marginal product curve crosses the horizontal axis of the graph at the point of maximum total product. This happens because adding a worker in a manner that slows production and decreases total output implies a negative marginal product for that worker.

The average product and marginal product curves are closely related. *When the marginal product is greater than the average product, the average product is increasing.* This is the case for labor inputs up to 4 in Figure 6.2(b). If the output of an additional worker is greater than the average output of each existing worker (i.e., the marginal product is greater than the average product), then adding the worker causes average output to rise. In Table 6.2, two workers produce 30 units of output, for an average product of 15 units per worker. Adding a third worker increases output by 30 units (to 60), which raises the average product from 15 to 20.

Similarly, *when the marginal product is less than the average product, the average product is decreasing.* This is the case when the labor input is greater than 4 in Figure 6.2(b). In Table 6.2, six workers produce 108 units of output, so that the average product is 18. Adding a seventh worker contributes a marginal product of only 4 units (less than the average product), reducing the average product to 16.

We have seen that the marginal product is above the average product when the average product is increasing, and below the average product when the average product is decreasing. It follows, therefore, that the marginal product must equal the average product when the average product reaches its maximum. This happens at point *E* in Figure 6.2(b).

Why, in practice, should we expect the marginal product curve to rise and then fall? Think of a television assembly plant. Fewer than ten workers might be insufficient to operate the assembly line at all. Ten to fifteen workers might be able to run the assembly line, but not very efficiently. Adding a few more workers might allow the assembly line to operate much more efficiently, so the marginal product of those workers would be very high. This added efficiency might start to diminish once there were more than 20 workers. The marginal product of the twenty-second worker, for example, might still be very high (and above the average product), but not as high as the marginal product of the nineteenth or twentieth worker. The marginal product of the twenty-fifth worker might be lower still, perhaps equal to the average product. With 30 workers, adding one more worker would yield more output, but not very much more (so that the marginal product, while positive, would be below the average product). Once there were more than 40 workers, additional workers would simply get in each other's way and actually reduce output (so that the marginal product would be negative).

The Average Product of Labor Curve

The geometric relationship between the total product and the average and marginal product curves is shown in Figure 6.2(a). The average product of labor is the total product divided by the quantity of labor input. At *B*, for example, the

average product is equal to the output of 60 divided by the input of 3, or 20 units of output per unit of labor input. This ratio, however, is exactly the slope of the line running from the origin to *B* in Figure 6.2(a). In general, *the average product of labor is given by the slope of the line drawn from the origin to the corresponding point on the total product curve.*

The Marginal Product of Labor Curve

The marginal product of labor is the change in the total product resulting from an increase of one unit of labor. At *A*, for example, the marginal product is 20 because the tangent to the total product curve has a slope of 20. In general, *the marginal product of labor at a point is given by the slope of the total product at that point.* We can see in Figure 6.2(a) that the marginal product of labor increases initially, peaks at an input of 3, and then declines as we move up the total product curve to *C* and *D*. At *D*, when total output is maximized, the slope of the tangent to the total product curve is 0, as is the marginal product. Beyond that point, the marginal product becomes negative.

The Relationship Between the Average and Marginal Products

Note the graphical relationship between average and marginal products in Figure 6.2(a). At *B*, the marginal product of labor (the slope of the tangent to the total product curve at *B*—not shown explicitly) is greater than the average product (dashed line *OB*). As a result, the average product of labor increases as we move from *B* to *C*. At *C*, the average and marginal products of labor are equal: While the average product is the slope of the line from the origin *OC*, the marginal product is the tangent to the total product curve at *C* (note the equality of the average and marginal products at point *E* in Figure 6.2(b)). Finally, as we move beyond *C* toward *D*, the marginal product falls below the average product; you can check that the slope of the tangent to the total product curve at any point between *C* and *D* is lower than the slope of the line from the origin.

The Law of Diminishing Marginal Returns

A diminishing marginal product of labor (and a diminishing marginal product of other inputs) holds for most production processes. The **law of diminishing marginal returns** states that as the use of an input increases in equal increments (with other inputs fixed), a point will eventually be reached at which the resulting additions to output decrease. When the labor input is small (and capital is fixed), extra labor adds considerably to output, often because workers are allowed to devote themselves to specialized tasks. Eventually, however, the law of diminishing marginal returns applies: When there are too many workers, some workers become ineffective and the marginal product of labor falls.

The law of diminishing marginal returns usually applies to the short run when at least one input is fixed. However, it can also apply to the long run. Even though inputs are variable in the long run, a manager may still want to analyze production choices for which one or more inputs are unchanged. Suppose, for example, that only two plant sizes are feasible and that management must decide which to build. In that case, management would want to know when diminishing marginal returns will set in for each of the two options.

Do not confuse the law of diminishing marginal returns with possible changes in the *quality* of labor as labor inputs are increased (as would likely occur, for example, if the most highly qualified laborers are hired first and the least qualified last). In our analysis of production, we have assumed that all

law of diminishing marginal returns Principle that as the use of an input increases with other inputs fixed, the resulting additions to output will eventually decrease.

labor inputs are of equal quality; diminishing marginal returns results from limitations on the use of other fixed inputs (e.g., machinery), not from declines in worker quality. In addition, do not confuse diminishing marginal returns with *negative* returns. The law of diminishing marginal returns describes a *declining* marginal product but not necessarily a negative one.

The law of diminishing marginal returns applies to a given production technology. Over time, however, inventions and other improvements in technology may allow the entire total product curve in Figure 6.2(a) to shift upward, so that more output can be produced with the same inputs. Figure 6.3 illustrates this principle. Initially the output curve is given by O_1 , but improvements in technology may allow the curve to shift upward, first to O_2 , and later to O_3 .

Suppose, for example, that over time, as labor is increased in agricultural production, technological improvements are being made. These improvements might include genetically engineered pesticide-resistant seeds, more powerful and effective fertilizers, and better farm equipment. As a result, output changes from A (with an input of 6 on curve O_1) to B (with an input of 7 on curve O_2) to C (with an input of 8 on curve O_3).

The move from A to B to C relates an increase in labor input to an increase in output and makes it appear that there is not diminishing marginal returns when in fact there is. Indeed, the shifting of the total product curve suggests that there may not be any negative long-run implications for economic growth. In fact, as we can see in Example 6.1, the failure to account for long-run improvements in technology led British economist Thomas Malthus wrongly to predict dire consequences from continued population growth.

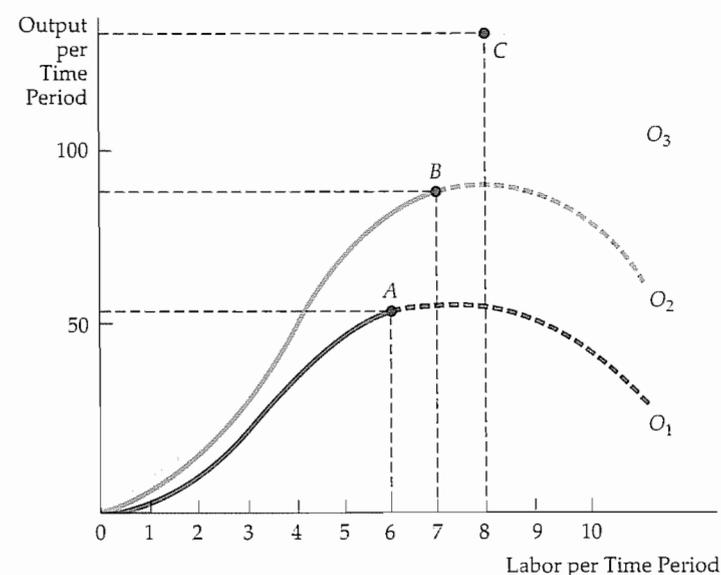


FIGURE 6.3 The Effect of Technological Improvement

Labor productivity (output per unit of labor) can increase if there are improvements in the technology, even though any given production process exhibits diminishing returns to labor. As we move from point A on curve O_1 to B on curve O_2 to C on curve O_3 over time, labor productivity increases.

EXAMPLE 6.1 Malthus and the Food Crisis

The law of diminishing marginal returns was central to the thinking of political economist Thomas Malthus (1766–1834).¹ Malthus believed that the limited amount of land on the globe would not be able to supply enough food as population grew and more laborers began to farm the land. Eventually as both the marginal and average productivity of labor fell and there were more mouths to feed, mass hunger and starvation would result. Fortunately, Malthus was wrong (although he was right about the diminishing marginal returns to labor).

Over the past century, technological improvements have dramatically altered the production of food in most countries (including developing countries, such as India). As a result, the average product of labor and total food output have increased. These improvements include new high-yielding, disease-resistant strains of seeds, better fertilizers, and better harvesting equipment. As Table 6.3 shows, overall food consumption throughout the world has outpaced population growth more or less continually since the end of World War II.² This increase in world agricultural productivity is also illustrated in Figure 6.4, which shows average cereal yields from 1970 through 1998, along with a world price index for food.³ Note that cereal yields have increased steadily over the period. Because growth in agricultural productivity led to increases in food supplies that outstripped the growth in demand, prices, apart from a temporary increase in the early 1970s, have been declining.

Some of the increase in food production has been due to small increases in the amount of land devoted to farming. From 1961 to 1975, for example, the percentage of land devoted to agriculture increased from 32.9 percent to 33.3 percent in Africa, from 19.6 percent to 22.4 percent in Latin America, and from

TABLE 6.3 Index of World Food Consumption Per Capita

YEAR	INDEX
1948–1952	100
1960	115
1970	123
1980	128
1990	137
1995	135
1998	140

¹ Thomas Malthus, *Essay on the Principle of Population*, 1798.

² All but the data for 1990, 1995, and 1998 appear as Table 4.1 in Julian Simon, *The Ultimate Resource* (Princeton: Princeton University Press, 1981). The original source for all the data is the UN Food and Agriculture Organization, *Production Yearbook*, and *World Agricultural Situation*.

³ Data are from the UN Food and Agriculture Organization and the World Bank. See also <http://apps.fao.org> (select Agriculture, then under "Data Collection," select Crops Primary).

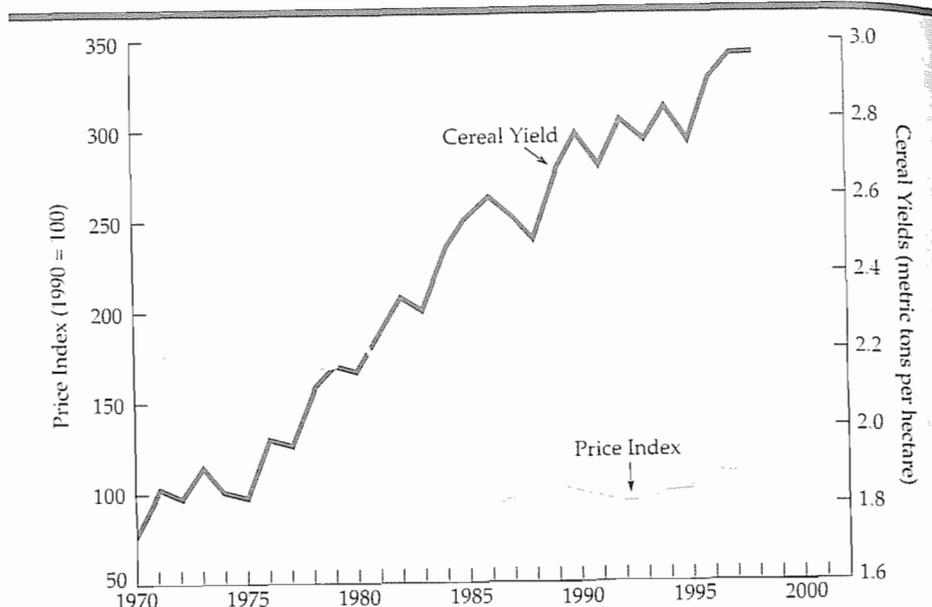


FIGURE 6.4 Cereal Yields and the World Price of Food

Cereal yields have increased steadily. The average world price of food increased temporarily in the early 1970s, but has declined since.

21.9 percent to 22.6 percent in the Far East.⁴ During the same period, however, the percentage of land devoted to agriculture fell from 26.1 percent to 25.5 percent in North America, and from 46.3 percent to 43.7 percent in Western Europe. It follows, therefore, that most of the improvement in food output is due to improved technology, not to increases in land used for agriculture.

Hunger remains a severe problem in some areas, such as the Sahel region of Africa, in part because of the low productivity of labor there. Although other countries produce an agricultural surplus, mass hunger still occurs because of the difficulty of redistributing foods from more to less productive regions of the world, and because of the low incomes of those less productive regions.

Labor Productivity

Although this is a textbook in microeconomics, many of the concepts developed here provide a foundation for macroeconomic analysis. Macroeconomists are particularly concerned with **labor productivity**—the average product of labor for an entire industry or for the economy as a whole. In this subsection we discuss labor productivity in the United States and in a number of foreign countries. This topic is interesting in its own right but will also help to illustrate one of the links between micro- and macroeconomics.

Because the average product measures output per unit of labor input, it is relatively easy to measure (total labor input and total output are the only pieces of information you need). Labor productivity can provide useful comparisons

across industries and for one industry over a long period. But labor productivity is especially important because it determines the real *standard of living* that a country can achieve for its citizens.

Productivity and the Standard of Living There is a simple link between labor productivity and the standard of living. In any particular year, the aggregate value of goods and services produced by an economy is equal to the payments made to all factors of production, including wages, rental payments to capital, and profit to firms. But consumers ultimately receive these factor payments, in the form of wages, salaries, dividends, or interest payments. As a result, consumers in the aggregate can increase their rate of consumption in the long run only by increasing the total amount they produce.

Understanding the causes of productivity growth is an important area of research in economics. We do know that one of the most important sources of growth in labor productivity is growth in the **stock of capital**—i.e., the total amount of capital available for use in production. Because an increase in capital means more and better machinery, each worker can produce more output for each hour worked. Another important source of growth in labor productivity is **technological change**—i.e., the development of new technologies that allow labor (and other factors of production) to be used more effectively and to produce new and higher-quality goods.

As Example 6.2 shows, levels of labor productivity have differed considerably across countries, and so too have rates of growth of productivity. Understanding these differences is important, given the central role that productivity has in affecting our standards of living.

stock of capital Total amount of capital available for use in production.

technological change Development of new technologies allowing factors of production to be used more effectively.

EXAMPLE 6.2 Labor Productivity and the Standard of Living

Will the standard of living in the United States, Europe, and Japan continue to improve, or will these economies barely keep future generations from being worse off than they are today? Because the real incomes of consumers in these countries increase only as fast as productivity does, the answer depends on the labor productivity of workers.

As Table 6.4 shows, the level of output per person in the United States in 1997 was higher than in other industrial countries. But two patterns over the post-World War II period have been disturbing for Americans. First, productivity in the United States has grown less rapidly than productivity in most

TABLE 6.4 Labor Productivity in Developed Countries

	FRANCE	GERMANY	JAPAN	UNITED KINGDOM	UNITED STATES
<i>Output per Employed Person (1997)</i>					
	\$54,507	\$55,644	\$46,048	\$42,630	\$60,916
<i>Annual Rate of Growth of Labor Productivity (%)</i>					
<i>Years</i>					
1960–1973	4.75	4.04	8.30	2.89	2.36
1974–1986	2.10	1.85	2.50	1.69	0.71
1987–1997	1.48	2.00	1.94	1.02	1.09

labor productivity Average product of labor for an entire industry or for the economy as a whole.

⁴ See Simon, *The Ultimate Resource*, p. 83.

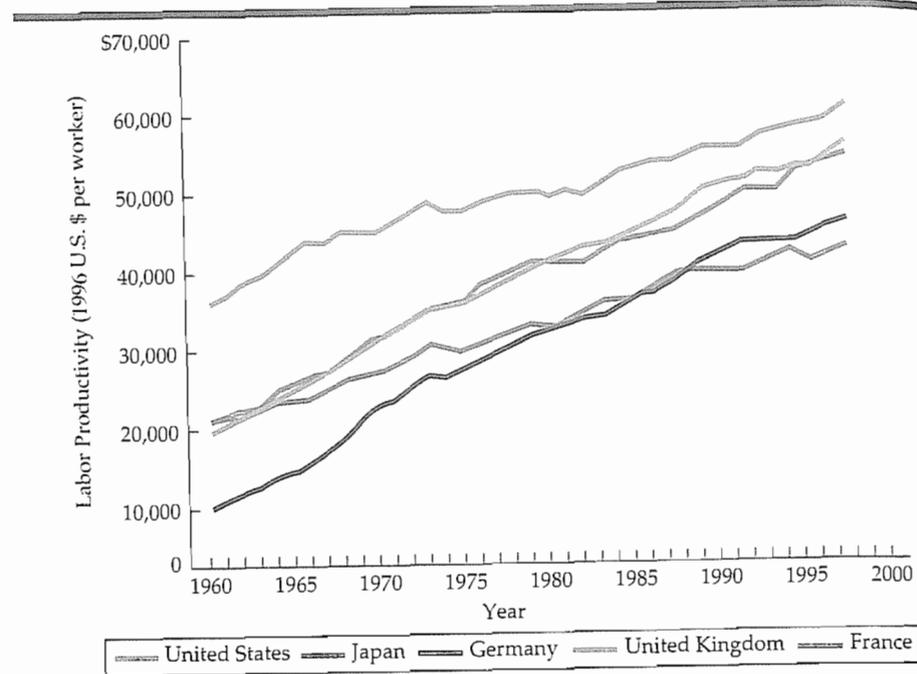


FIGURE 6.5 Labor Productivity in Five Countries

During the 1960s and 1970s, productivity growth in the United States was lower than in Germany, France, the United Kingdom, and Japan, although the level of productivity was higher. In the 1980s and 1990s, productivity growth slowed in all of these countries.

other developed nations. Second, productivity growth during 1974–1997 was much lower in all developed countries than it had been in the past. Both of these patterns can be seen in the table and in Figure 6.5.⁵ The figure shows productivity, measured in 1997 U.S. dollars per worker, for the United States and for four other countries. Observe that in 1960 labor productivity in the United States was more than three times labor productivity in Japan and about twice as great as labor productivity in Germany, France, and the United Kingdom. However, by 1997 the differences had narrowed considerably.

Throughout most of the 1960–1997 period, Japan had the highest rate of productivity growth, followed by Germany and France. U.S. productivity growth was the lowest, even somewhat lower than that of the United Kingdom. This is partly due to differences in rates of investment and growth in the stock of capital in each country. The greatest capital growth during the postwar period was in Japan and France, which were rebuilt substantially after World War II. To some extent, therefore, the lower rate of growth of productivity in the United States, when compared to that of Japan, France, and Germany, is the result of these countries catching up after the war.

⁵ Recent growth numbers are based on data from *Industrial Policy in OECD Countries, Annual Review*, and *International Comparisons of Manufacturing Productivity and Unit Labor Cost Trends*, U.S. Bureau of Labor Statistics (1998), www.stats.bls.gov/flsdata.htm.

Productivity growth is also tied to the natural resource sector of the economy. As oil and other resources began to be depleted, output per worker fell. Environmental regulations (e.g., the need to restore land to its original condition after strip-mining for coal) magnified this effect as the public became more concerned with the importance of cleaner air and water.

Observe from Table 6.4 that productivity growth in the United States has increased in recent years. Economists have debated whether this is a short-term aberration or the beginning of a long-term trend. Some economists believe that rapid technological change during the 1990s, and in particular the computer revolution, has created new possibilities for productivity growth. If this optimistic view is correct, we will see continued high rates of productivity growth in the coming years.⁶

6.4 Production with Two Variable Inputs

Now that we have seen the relationship between production and productivity, let's turn to production in the long run, where both capital and labor inputs are variable. The firm can now produce its output in a variety of ways by combining different amounts of labor and capital. We will use isoquants to analyze and compare these different ways of producing.

Recall that an isoquant describes all combinations of inputs that yield the same level of output. The isoquants shown in Figure 6.6 are reproduced from Figure 6.1; they all slope downward because both labor and capital have positive marginal products. More of either input increases output; thus if output is to be kept constant as more of one input is used, less of the other input must be used.

Diminishing Marginal Returns

Even though both labor and capital are variable in the long run, it is useful for a firm that is choosing the optimal mix of inputs to ask what happens to output as each of the inputs is increased, with the other input held fixed. The outcome of this exercise is described in Figure 6.6, which reflects diminishing marginal returns to both labor and capital. We can see why there is diminishing marginal returns to labor by drawing a horizontal line at a particular level of capital—say, 3. Reading the levels of output from each isoquant as labor is increased, we note that each additional unit of labor generates less and less additional output. For example, when labor is increased from 1 unit to 2 (from A to B), output increases by 20 (from 55 to 75). However, when labor is increased by an additional unit (from B to C), output increases by only 15 (from 75 to 90). Thus there is diminishing marginal returns to labor both in the long and short run. Because adding one factor while holding the other factor constant eventually leads to lower and lower incremental output, the isoquant must become steeper as more capital is added in place of labor and flatter when labor is added in place of capital.

⁶ For more information on labor productivity and standard of living, go to <http://stats.bls.gov/flsdata.htm>. Under "International Comparisons of Productivity, Unit Labor Costs, and GDP per Capita," click on: Unpublished Comparative Real Gross Domestic Product per Capita and per Employed Person, Fourteen Countries, 1960–1996.

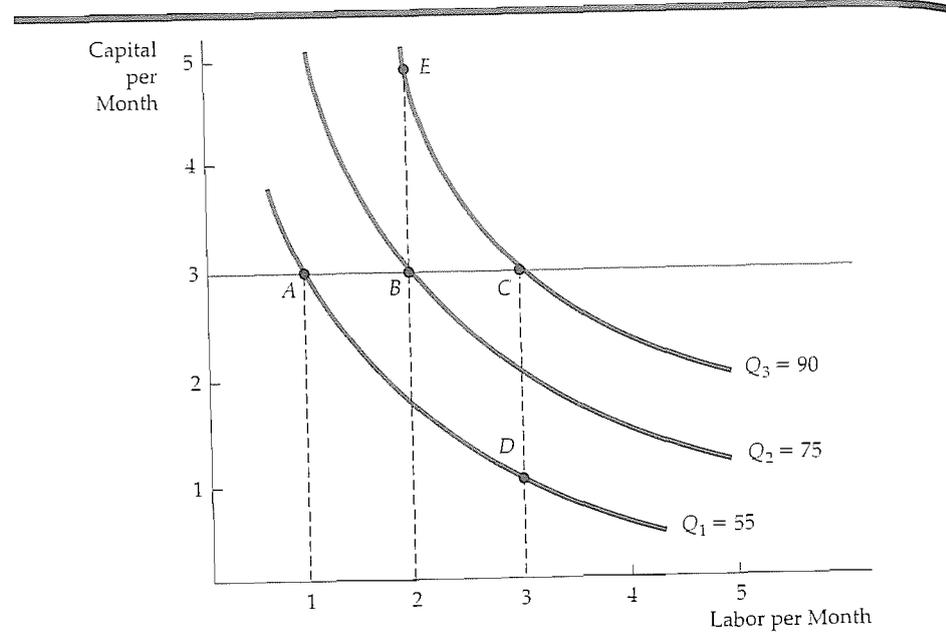


FIGURE 6.6 The Shape of Isoquants

When both labor and capital are variable, both factors of production can exhibit diminishing marginal returns. As we move from A to C, there is diminishing returns to labor, and as we move from D to C, there is diminishing returns to capital.

There is also diminishing marginal returns to capital. With labor fixed, the marginal product of capital decreases as capital is increased. For example, when capital is increased from 1 to 2 and labor is held constant at 3, the marginal product of capital is initially 20 (75-55) but falls to 15 (90-75) when capital is increased from 2 to 3.

Substitution Among Inputs

With two inputs that can be varied, a manager will want to consider substituting one input for another. The slope of each isoquant indicates how the quantity of one input can be traded off against the quantity of the other, while output is held constant. When the negative sign is removed, we call the slope the **marginal rate of technical substitution (MRTS)**. The *marginal rate of technical substitution of labor for capital* is the amount by which the input of capital can be reduced when one extra unit of labor is used, so that output remains constant. This is analogous to the marginal rate of substitution (MRS) in consumer theory. Recall from Section 3.1 that the MRS describes how consumers substitute among two goods while holding the level of satisfaction constant. Like the MRS, the MRTS is always measured as a positive quantity:

$$\begin{aligned} \text{MRTS} &= - \text{Change in capital input} / \text{change in labor input} \\ &= - \Delta K / \Delta L \text{ (for a fixed level of } Q) \end{aligned}$$

where ΔK and ΔL are small changes in capital and labor along an isoquant.

marginal rate of technical substitution (MRTS) Amount by which the quantity of one input can be reduced when one extra unit of another input is used, so that output remains constant.

In §3.1, we explain that the marginal rate of substitution is the maximum amount of one good that the consumer is willing to give up to obtain one unit of another good.

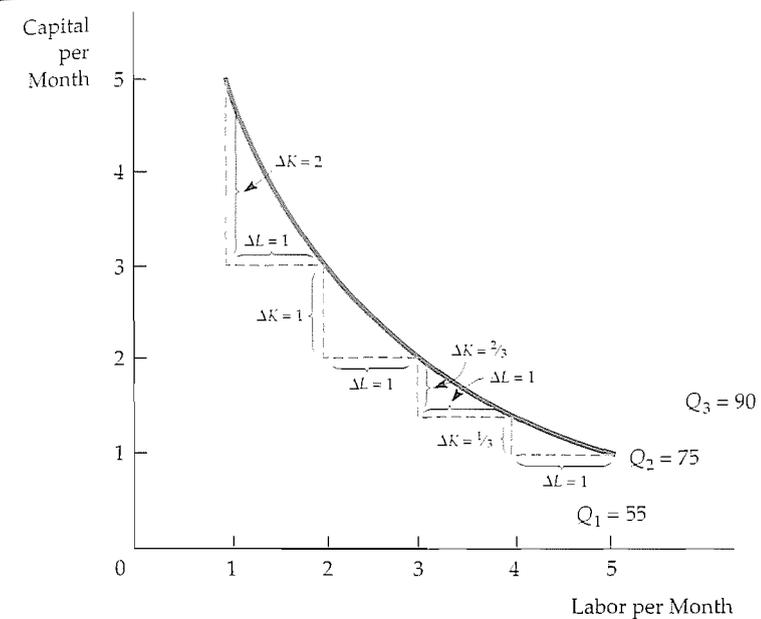


FIGURE 6.7 Marginal Rate of Technical Substitution

Like indifference curves, isoquants are downward sloping and convex. The slope of the isoquant at any point measures the marginal rate of technical substitution—the ability of the firm to replace capital with labor while maintaining the same level of output. On isoquant Q_2 , the MRTS falls from 2 to 1 to $2/3$ to $1/3$.

In Figure 6.7 the MRTS is equal to 2 when labor increases from 1 unit to 2 and output is fixed at 75. However, the MRTS falls to 1 when labor is increased from 2 units to 3, and then declines to $2/3$ and to $1/3$. Clearly, as more and more labor replaces capital, labor becomes less productive and capital becomes relatively more productive. Therefore we need less capital to keep output constant, and the isoquant becomes flatter.

Diminishing MRTS We assume that there is a *diminishing MRTS*. In other words, the MRTS falls as we move down along an isoquant. The mathematical implication is that isoquants, like indifference curves, are *convex*, or bowed inward. This is indeed the case for most production technologies. The diminishing MRTS tells us that the productivity of any one input is limited. As more and more labor is added to the production process in place of capital, the productivity of labor falls. Similarly, when more capital is added in place of labor, the productivity of capital falls. Production needs a balanced mix of both inputs.

As our discussion has just suggested, the MRTS is closely related to the marginal products of labor MP_L and capital MP_K . To see how, imagine adding some labor and reducing the amount of capital sufficient to keep output constant. The additional output resulting from the increased labor input is equal to the additional output per unit of additional labor (the marginal product of labor) times the number of units of additional labor:

$$\text{Additional output from increased use of labor} = (MP_L)(\Delta L)$$

In §3.1, we explain that an indifference curve is convex if the marginal rate of substitution diminishes as we move down along the curve.

Similarly, the decrease in output resulting from the reduction in capital is the loss of output per unit reduction in capital (the marginal product of capital) times the number of units of capital reduction:

$$\text{Reduction in output from decreased use of capital} = (MP_K)(\Delta K)$$

Because we are keeping output constant by moving along an isoquant, the total change in output must be zero. Thus,

$$(MP_L)(\Delta L) + (MP_K)(\Delta K) = 0$$

Now, by rearranging terms we see that

$$(MP_L)/(MP_K) = -(\Delta K/\Delta L) = \text{MRTS} \quad (6.2)$$

Equation (6.2) tells us that *the marginal rate of technical substitution between two inputs is equal to the ratio of the marginal physical products of the inputs*. This formula will be useful when we look at the firm's cost-minimizing choice of inputs in Chapter 7.

Production Functions—Two Special Cases

Two extreme cases of production functions show the possible range of input substitution in the production process. In the first case, shown in Figure 6.8, inputs to production are *perfect substitutes* for one another. Here the MRTS is constant at all points on an isoquant. As a result, the same output (say Q_3) can be produced with mostly capital (at A), with mostly labor (at C), or with a balanced

In §3.1, we explain that two goods are perfect substitutes if the marginal rate of substitution of one for the other is a constant.

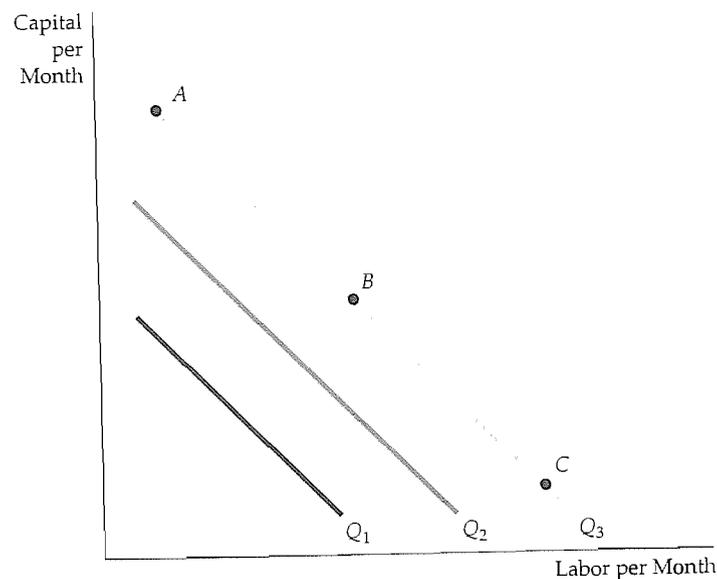


FIGURE 6.8 Isoquants When Inputs Are Perfect Substitutes

When the isoquants are straight lines, the MRTS is constant. Thus the rate at which capital and labor can be substituted for each other is the same no matter what level of inputs is being used. Points A, B, and C represent three different capital-labor combinations that generate the same output Q_3 .

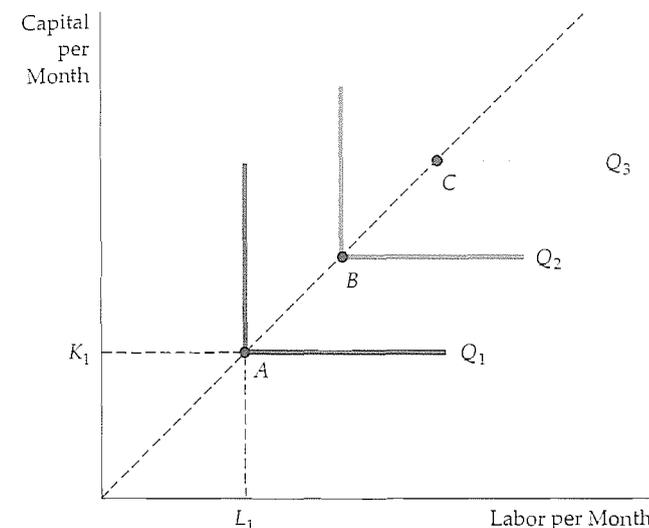


FIGURE 6.9 Fixed-Proportions Production Function

When the isoquants are L-shaped, only one combination of labor and capital can be used to produce a given output (as at point A on isoquant Q_1 , point B on isoquant Q_2 , and point C on isoquant Q_3). Adding more labor alone does not increase output, nor does adding more capital alone.

combination of both (at B). For example, musical instruments can be manufactured almost entirely with machine tools or with very few tools and highly skilled labor.

Figure 6.9 illustrates the opposite extreme, the **fixed-proportions production function**. In this case, it is impossible to make any substitution among inputs. Each level of output requires a specific combination of labor and capital: Additional output cannot be obtained unless more capital and labor are added in specific proportions. As a result, the isoquants are L-shaped just as indifference curves are L-shaped when two goods are perfect complements. An example is the reconstruction of concrete sidewalks using jackhammers. It takes one person to use a jackhammer—neither two people and one jackhammer nor one person and two jackhammers will increase production. As another example, suppose that a cereal company offers a new breakfast cereal, Nutty Oat Crunch, whose two inputs, not surprisingly, are oats and nuts. The secret formula for the cereal requires exactly one ounce of nuts for every four ounces of oats in every cereal serving. If the company were to purchase additional nuts but not additional oats, the output of cereal would remain unchanged, since the nuts must be combined with the oats in fixed proportions. Similarly, purchasing additional oats without additional nuts would also be unproductive.

In Figure 6.9 points A, B, and C represent technically efficient combinations of inputs. For example, to produce output Q_1 , a quantity of labor L_1 and capital K_1 can be used, as at A. If capital stays fixed at K_1 , adding more labor does not change output. Nor does adding capital with labor fixed at L_1 . Thus on the vertical and the horizontal segments of the L-shaped isoquants, either the marginal product of capital or the marginal product of labor is zero. Higher output results only when both labor and capital are added, as in the move from input combination A to input combination B.

fixed-proportions production function Production function with L-shaped isoquants, so that only one combination of labor and capital can be used to produce each level of output.

In §3.1, we explain that two goods are perfect complements when the indifference curves for the goods are shaped as right angles.

The fixed-proportions production function describes situations in which methods of production are limited. For example, the production of a television show might involve a certain mix of capital (camera and sound equipment, etc.) and labor (producer, director, actors, etc.). To make more television shows, all inputs to production must be increased proportionally. In particular, it would be difficult to increase capital inputs at the expense of labor, because actors are necessary inputs to production (except perhaps for animated films). Likewise, it would be difficult to substitute labor for capital, because filmmaking today requires sophisticated film equipment.

EXAMPLE 6.3 A Production Function for Wheat

Crops can be produced using different methods. Food grown on large farms in the United States is usually produced with a *capital-intensive technology*, which involves substantial investments in capital, such as buildings and equipment, and relatively little input of labor. However, food can also be produced using very little capital (a hoe) and a lot of labor (several people with the patience and stamina to work the soil). One way to describe the agricultural production process is to show one isoquant (or more) that describes the combination of inputs that generates a given level of output (or several output levels). The description that follows comes from a production function for wheat that was estimated statistically.⁷

Figure 6.10 shows one isoquant, associated with the production function, corresponding to an output of 13,800 bushels of wheat per year. The manager of the farm can use this isoquant to decide whether it is profitable to hire more labor or use more machinery. Assume the farm is currently operating at *A*, with a labor input *L* of 500 hours and a capital input *K* of 100 machine hours. The manager decides to experiment by using only 90 hours of machine time. To produce the same crop per year, he finds that he needs to replace this machine time by adding 260 hours of labor.

The results of this experiment tell the manager about the shape of the wheat production isoquant. When he compares points *A* (where $L = 500$ and $K = 100$) and *B* (where $L = 760$ and $K = 90$) in Figure 6.10, both of which are on the same isoquant, the manager finds that the marginal rate of technical substitution is equal to 0.04 ($-\Delta K/\Delta L = -(-10)/260 = .04$).

The MRTS tells the manager the nature of the trade-off involved in adding labor and reducing the use of farm machinery. Because the MRTS is substantially less than 1 in value, the manager knows that when the wage of a laborer is equal to the cost of running a machine, he ought to use more capital. (At his current level of production, he needs 260 units of labor to substitute for 10 units of capital.) In fact, he knows that unless labor is much less expensive than the use of a machine, his production process ought to become more capital-intensive.

The decision about how many laborers to hire and machines to use cannot be fully resolved until we discuss the costs of production in the next chapter. However, this example illustrates how knowledge about production isoquants and the marginal rate of technical substitution can help a manager. It also suggests why most farms in the United States and Canada, where labor is rela-

⁷ The food production function on which this example is based is given by the equation $Q = 100(K^{\frac{1}{4}}L^{\frac{3}{4}})$, where Q is the rate of output in bushels of food per year, K is the quantity of machines in use per year, and L is the number of hours of labor per year.

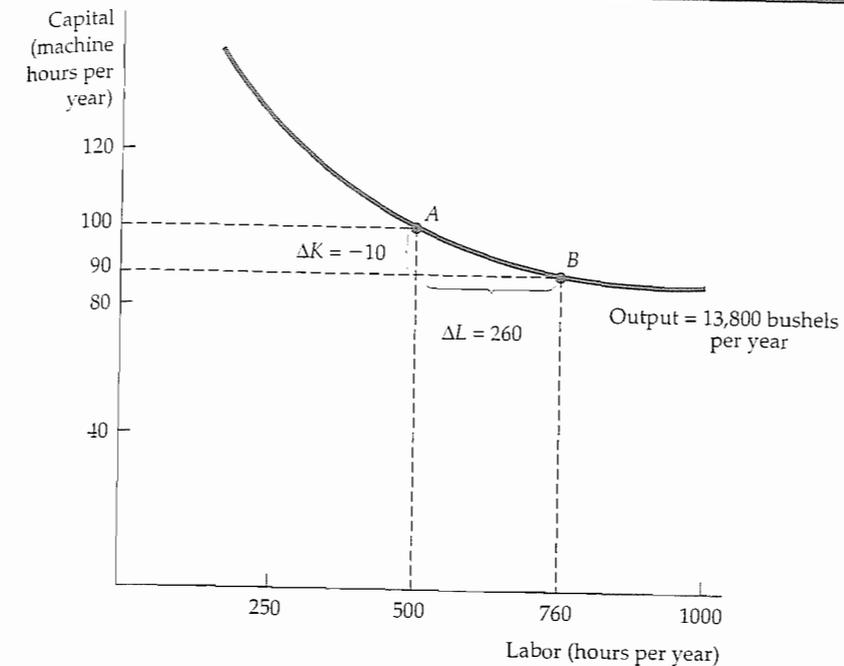


FIGURE 6.10 Isoquant Describing the Production of Wheat

A wheat output of 13,800 bushels per year can be produced with different combinations of labor and capital. The more capital-intensive production process is shown as point *A*, the more labor-intensive process as point *B*. The marginal rate of technical substitution between *A* and *B* is $10/260 = 0.04$.

tively expensive, operate in the range of production in which the MRTS is relatively high (with a high capital-to-labor ratio), whereas farms in developing countries, in which labor is cheap, operate with a lower MRTS (and a lower capital-to-labor ratio).⁸ The exact labor/capital combination to use depends on input prices, a subject we discuss in Chapter 7.

6.5 Returns to Scale

Our analysis of input substitution in the production process has shown us what happens when a firm substitutes one input for another while keeping output constant. However, in the long run, with all inputs variable, the firm must also consider the best way to increase output. One way to do so is to change the *scale* of the operation by increasing *all of the inputs to production in proportion*. If it takes one farmer working with one harvesting machine on one acre of land to produce

⁸ With the production function given in footnote 7, it is not difficult (using calculus) to show that the marginal rate of technical substitution is given by $MRTS = (MP_L/MP_K) = (1/4)(K/L)$. Thus the MRTS decreases as the capital-to-labor ratio falls. For an interesting study of agricultural production in Israel, see Richard E. Just, David Zilberman, and Eithan Hochman, "Estimation of Multicrop Production Functions," *American Journal of Agricultural Economics* 65 (1983): 770–80.

returns to scale Rate at which output increases as inputs are increased proportionately.

increasing returns to scale Output more than doubles when all inputs are doubled.

constant returns to scale Output doubles when all inputs are doubled.

decreasing returns to scale Output less than doubles when all inputs are doubled.

100 bushels of wheat, what will happen to output if we put two farmers to work with two machines on two acres of land? Output will almost certainly increase, but will it double, more than double, or less than double? **Returns to scale** is the rate at which output increases as inputs are increased proportionately. We will examine three different cases: increasing, constant, and decreasing returns to scale.

Increasing Returns to Scale

If output more than doubles when inputs are doubled, there are **increasing returns to scale**. This might arise because the larger scale of operation allows managers and workers to specialize in their tasks and to make use of more sophisticated, large-scale factories and equipment. The automobile assembly line is a famous example of increasing returns.

The prospect of increasing returns to scale is an important issue from a public policy perspective. If there are increasing returns, then it is economically advantageous to have one large firm producing (at relatively low cost) rather than to have many small firms (at relatively high cost). Because this large firm can control the price that it sets, it may need to be regulated. For example, increasing returns in the provision of electricity is one reason why we have large, regulated power companies.

Constant Returns to Scale

A second possibility with respect to the scale of production is that output may double when inputs are doubled. In this case, we say there are **constant returns to scale**. With constant returns to scale, the size of the firm's operation does not affect the productivity of its factors—one plant using a particular production process can easily be replicated, so that two plants produce twice as much output. For example, a large travel agency might provide the same service per client and use the same ratio of capital (office space) and labor (travel agents) as a small agency that services fewer clients.

Decreasing Returns to Scale

Finally, output may less than double when all inputs double. This case of **decreasing returns to scale** applies to some firms with large-scale operations. Eventually, difficulties in organizing and running a large-scale operation may lead to decreased productivity of both labor and capital. Communication between workers and managers can become difficult to monitor as the workplace becomes more impersonal. Thus the decreasing-returns case is likely to be associated with the problems of coordinating tasks and maintaining a useful line of communication between management and workers.

Describing Returns to Scale

The presence or absence of returns to scale is seen graphically in the two parts of Figure 6.11. The line OA from the origin in each panel describes a production process in which labor and capital are used as inputs to produce various levels of output in the ratio of 5 hours of labor to 2 hours of machine time. In Figure 6.11(a), the firm's production function exhibits constant returns to scale. When 5

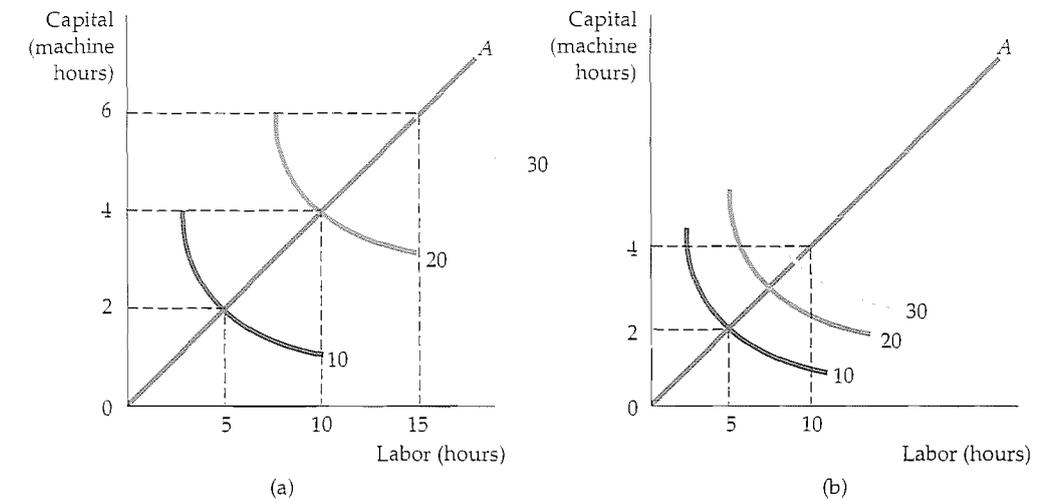


FIGURE 6.11 Returns to Scale

When a firm's production process exhibits constant returns to scale as shown by a movement along ray OA in part (a), the isoquants are equally spaced as output increases proportionally. However, when there are increasing returns to scale as shown in (b), the isoquants move closer together as inputs are increased along the ray.

hours of labor and 2 hours of machine time are used, an output of 10 units is produced. When both inputs double, output doubles from 10 to 20 units; when both inputs triple, output triples, from 10 to 30 units. Put differently, twice as much of both inputs is needed to produce 20 units, and three times as much is needed to produce 30 units.

In Figure 6.11(b), the firm's production function exhibits increasing returns to scale. Now the isoquants become closer together as we move away from the origin along OA . As a result, less than twice the amount of both inputs is needed to increase production from 10 units to 20; substantially less than three times the inputs are needed to produce 30 units. The reverse would be true if the production function exhibited decreasing returns to scale (not shown here). With decreasing returns, the isoquants become increasingly distant from one another as output levels proportionally increase.

Returns to scale vary considerably across firms and industries. Other things being equal, the greater the returns to scale, the larger firms in an industry are likely to be. Because manufacturing involves large investments in capital equipment, manufacturing industries are more likely to have increasing returns to scale than service-oriented industries. Services are more labor-intensive and can usually be provided as efficiently in small quantities as they can on a large scale.

EXAMPLE 6.4 Returns to Scale in the Carpet Industry

The carpet industry in the United States centers around the town of Dalton in northern Georgia. From a relatively small industry with many small firms in the first half of the twentieth century, it grew rapidly and became a

TABLE 6.5 The U.S. Carpet Industry

CARPET SHIPMENTS, 1996 (MILLIONS OF DOLLARS PER YEAR)			
1. Shaw Industries	3,202	6. World Carpets	475
2. Mohawk Industries	1,795	7. Burlington Industries	450
3. Beaulieu of America	1,006	8. Collins & Aikman	418
4. Interface Flooring	820	9. Masland Industries	380
5. Queen Carpet	775	10. Dixie Yarns	280

major industry with a large number of firms of all sizes. For example, the top ten carpet manufacturers, ranked by shipments in millions of dollars in 1996, are shown in Table 6.5.⁹

Currently, there are three relatively large manufacturers (Shaw, Mohawk, and Beaulieu), along with a number of smaller producers. There are also many carpet retailers, wholesale distributors, carpet buying groups, and national retail carpet chains. The carpet industry has grown rapidly for several reasons. Consumer demand for wool, nylon, and polypropylene carpets in commercial and residential uses has skyrocketed. In addition, innovations such as the introduction of larger, faster, and more efficient carpet-tufting machines have reduced costs and greatly increased carpet production. Along with the increase in production, innovation and competition have worked together to reduce real carpet prices.

To what extent, if any, can the growth of the carpet industry be explained by the presence of returns to scale? There have certainly been substantial improvements in the processing of key production inputs (such as stain-resistant yarn) and in the distribution of carpets to retailers and consumers. But what about the production of carpets? Carpet production is capital intensive—manufacturing plants require heavy investments in high-speed tufting machines that turn various types of yarn into carpet, as well as machines that put the backings onto the carpets, cut the carpets into appropriate sizes, and package, label, and distribute them.

Overall, physical capital (including plant and equipment) accounts for about 77 percent of a typical carpet manufacturer's costs, while labor accounts for the remaining 23 percent. Over time, the major carpet manufacturers have increased the scale of their operations by putting larger and more efficient tufting machines into larger plants. At the same time, the use of labor in these plants has also increased significantly. The result? Proportional increases in inputs have resulted in a more than proportional increase in output for these larger plants. For example, a doubling of capital and labor inputs might lead to a 110-percent increase in output. This pattern has not, however, been uniform across the industry. Most smaller carpet manufacturers have found that small changes in scale have little or no effect on output; i.e., small proportional increases in inputs have only increased output proportionally.

⁹ Frank O'Neill, "The Focus 100," *Focus* (May 1997): 20.

We can therefore characterize the carpet industry as one in which there are constant returns to scale for relatively small plants but increasing returns to scale for larger plants. These increasing returns, however, are limited, and we can expect that if plant size were increased further, there would eventually be decreasing returns to scale.

SUMMARY

1. A *production function* describes the maximum output a firm can produce for each specified combination of inputs.
2. An *isoquant* is a curve that shows all combinations of inputs that yield a given level of output. A firm's production function can be represented by a series of isoquants associated with different levels of output.
3. In the short run, one or more inputs to the production process are fixed. In the long run, all inputs are potentially variable.
4. Production with one variable input, labor, can be usefully described in terms of the *average product of labor* (which measures output per unit of labor input), and the *marginal product of labor* (which measures the additional output as labor is increased by 1 unit).
5. According to the *law of diminishing marginal returns*, when one or more inputs are fixed, a variable input (usually labor) is likely to have a marginal product that eventually diminishes as the level of input increases.
6. Isoquants always slope downward because the marginal product of all inputs is positive. The shape of each isoquant can be described by the marginal rate of technical substitution at each point on the isoquant. The *marginal rate of technical substitution of labor for capital* (MRTS) is the amount by which the input of capital can be reduced when one extra unit of labor is used so that output remains constant.
7. The standard of living that a country can attain for its citizens is closely related to its level of labor productivity. Decreases in the rate of productivity growth in developed countries are due in part to the lack of growth of capital investment.
8. The possibilities for substitution among inputs in the production process range from a production function in which inputs are *perfect substitutes* to one in which the proportions of inputs to be used are fixed (a *fixed-proportions production function*).
9. In long-run analysis, we tend to focus on the firm's choice of its scale or size of operation. Constant returns to scale means that doubling all inputs leads to doubling output. Increasing returns to scale occurs when output more than doubles when inputs are doubled; decreasing returns to scale applies when output less than doubles.

QUESTIONS FOR REVIEW

1. What is a production function? How does a long-run production function differ from a short-run production function?
2. Why is the marginal product of labor likely to increase and then decline in the short run?
3. Diminishing returns to a single factor of production and constant returns to scale are not inconsistent. Discuss.
4. You are an employer seeking to fill a vacant position on an assembly line. Are you more concerned with the average product of labor or the marginal product of labor for the last person hired? If you observe that your average product is just beginning to decline, should you hire any more workers? What does this situation imply about the marginal product of your last worker hired?
5. Faced with constantly changing conditions, why would a firm ever keep *any* factors fixed? What criteria determine whether a factor is fixed or variable?
6. How does the curvature of an isoquant relate to the marginal rate of technical substitution?

7. Can a firm have a production function that exhibits increasing returns to scale, constant returns to scale, and decreasing returns to scale at different scales of production as output increases? Discuss.
8. Give an example of a production process in which the short run involves a day or a week, and the long run any period longer than a week.

EXERCISES

1. Suppose a chair manufacturer is producing in the short run when equipment is fixed. The manufacturer knows that as the number of laborers used in the production process increases from 1 to 7, the number of chairs produced changes as follows: 10, 17, 22, 25, 26, 25, 23.
 - a. Calculate the average and marginal product of labor for this production function.
 - b. Does this production function exhibit diminishing returns to labor? Explain.
 - c. Explain intuitively what might cause the marginal product of labor to become negative.
2. Fill in the gaps in the table below.

QUANTITY OF VARIABLE INPUT	TOTAL OUTPUT	MARGINAL PRODUCT OF VARIABLE INPUT	AVERAGE PRODUCT OF VARIABLE INPUT
0	0	—	—
1	150		
2			200
3		200	
4	760		
5		150	
6			150
3. A political campaign manager must decide whether to emphasize television advertisements or letters to potential voters. Describe the production function for votes. How might information about this function (such as the shape of the isoquants) help the campaign manager to plan strategy?
4. A firm has a production process in which the inputs to production are perfectly substitutable in the long run. Can you tell whether the marginal rate of technical substitution is high or low, or is further information necessary? Discuss.
5. The marginal product of labor is known to be greater than the average product of labor at a given level of employment. Is the average product increasing or decreasing? Explain.
6. The marginal product of labor in the production of computer chips is 50 chips per hour. The marginal rate of technical substitution of hours of labor for hours of machine-capital is 1/4. What is the marginal product of capital?
7. Do the following production functions exhibit decreasing, constant, or increasing returns to scale?
 - a. $Q = .5KL$
 - b. $Q = 2K + 3L$
8. The production function for the personal computers of DISK, Inc., is given by $Q = 10K^3L^5$, where Q is the number of computers produced per day, K is hours of machine time, and L is hours of labor input. DISK's competitor, FLOPPY, Inc., is using the production function $Q = 10K^6L^4$.
 - a. If both companies use the same amounts of capital and labor, which will generate more output?
 - b. Assume that while capital is limited to 9 machine hours, labor is unlimited in supply. In which company is the marginal product of labor greater? Explain.
9. In Example 6.3, wheat is produced according to the production function $Q = 100(K^3L^2)$.
 - a. Beginning with a capital input of 4 and a labor input of 49, show that the marginal product of labor and the marginal product of capital are both decreasing.
 - b. Does this production function exhibit increasing, decreasing, or constant returns to scale?

CHAPTER 7

The Cost of Production

Chapter Outline

- 7.1 Measuring Cost: Which Costs Matter? 203
 - 7.2 Cost in the Short Run 208
 - 7.3 Cost in the Long Run 215
 - 7.4 Long-Run versus Short-Run Cost Curves 224
 - 7.5 Production with Two Outputs—Economies of Scope 229
 - *7.6 Dynamic Changes in Costs—The Learning Curve 232
 - *7.7 Estimating and Predicting Cost 237
- Appendix to Chapter 7:
Production and Cost Theory—A Mathematical Treatment 246

List of Examples

- 7.1 Choosing the Location for a New Law School Building 205
- 7.2 Sunk, Fixed, and Variable Costs: Computers, Software, and Pizzas 207
- 7.3 The Short-Run Cost of Aluminum Smelting 213
- 7.4 The Effect of Effluent Fees on Input Choices 220
- 7.5 Economies of Scope in the Trucking Industry 232
- 7.6 The Learning Curve in Practice 236
- 7.7 Cost Functions for Electric Power 240
- 7.8 A Cost Function for the Savings and Loan Industry 241

In the last chapter, we examined the firm's production technology—the relationship that shows how factor inputs can be transformed into outputs. Now we will see how the production technology, together with the prices of factor inputs, determine the firm's cost of production.

Given a firm's production technology, managers must decide *how* to produce. As we saw, inputs can be combined in different ways to yield the same amount of output. For example, one can produce a certain output with a lot of labor and very little capital, with very little labor and a lot of capital, or with some other combination of the two. In this chapter we see how the *optimal*—i.e., cost-minimizing—combination of inputs is chosen. We will also see how a firm's costs depend on its rate of output and show how these costs are likely to change over time.

We begin by explaining how *cost* is defined and measured, distinguishing between the concept of cost used by economists, who are concerned about the firm's future performance, and by accountants, who focus on the firm's financial statements. We then examine how the characteristics of the firm's production technology affect costs, both in the short run, when the firm can do little to change its capital stock, and in the long run, when the firm can change all its factor inputs.

We then show how the concept of returns to scale can be generalized to allow for *both* changes in the mix of inputs and the production of many different outputs. We also show how cost sometimes falls over time as managers and workers learn from experience and make the production process more efficient. Finally, we show how empirical information can be used to estimate cost functions and predict future costs.

7.1 Measuring Cost: Which Costs Matter?

Before we can analyze how a firm can minimize costs, we must clarify what we mean by *cost* in the first place and how we should measure it. What items, for example, should be included as part of a firm's cost? Cost obviously includes the wages a firm pays its workers and the rent it pays for office

space. But what if the firm already owns an office building and doesn't have to pay rent? How should we treat money that the firm spent two or three years ago (and can't recover) for equipment or for research and development? We'll answer questions such as these in the context of the economic decisions that managers make.

Economic Cost versus Accounting Cost

Economists often think of cost differently from financial accountants, who are usually concerned with reporting the past performance of the firm for external use, as in annual reports. Financial accountants tend to take a retrospective view of the firm's finances and operations because they must keep track of assets and liabilities and evaluate past performance. As a result, **accounting cost**—the cost that financial accountants measure—can include items that an economist would not include and would not include items that economists usually do include. For example, accounting cost includes actual expenses plus depreciation expenses for capital equipment, which are determined on the basis of the allowable tax treatment by the Internal Revenue Service.

Economists—and we hope managers—take a forward-looking view of the firm. They are concerned with the allocation of scarce resources. Therefore, they care about what cost is likely to be in the future and about ways in which the firm might be able to rearrange its resources to lower its costs and improve its profitability. As we will see, economists are therefore concerned with **economic cost**, which is the cost associated with forgone opportunities. The word *economic* tells us to distinguish between costs that the firm can control and those it cannot.

Opportunity Cost

Economists use the terms economic cost and *opportunity cost* synonymously. **Opportunity cost** is the cost associated with opportunities that are forgone by not putting the firm's resources to their highest-value use. For example, consider a firm that owns a building and therefore pays no rent for office space. Does this mean that the cost of office space is zero? While a financial accountant would treat this cost as zero, an economist would note that the firm could have earned rent on the office space by leasing it to another company. This forgone rent is the opportunity cost of utilizing the office space and should be included as part of the economic cost of doing business.

Accountants and economists both include actual monetary outlays, called *cash flows*, in their calculations. Cash flows include wages, salaries, and the cost of materials and property rentals; they are important because they involve direct payments to other firms and individuals. These costs are relevant for the economist because most monetary outlays, including wages and materials costs, represent money that could have usefully been spent elsewhere.

Let's take a look at how opportunity cost can make economic cost differ from accounting cost in the treatment of wages and economic depreciation. Consider an owner who manages her own retail store but chooses not to pay herself a salary. Although no monetary transaction has occurred (and thus no accounting cost is recorded), the business nonetheless incurs an opportunity cost because the owner could have earned a competitive salary by working elsewhere.

Likewise, accountants and economists often treat depreciation differently. When estimating the future profitability of a business, an economist or manager is concerned with the capital cost of plant and machinery. This cost involves not

accounting cost Actual expenses plus depreciation charges for capital equipment.

economic cost Cost to a firm of utilizing economic resources in production, including opportunity cost.

opportunity cost Cost associated with opportunities that are forgone when a firm's resources are not put to their highest-value use.

only the monetary outlay for buying and then running the machinery, but also the cost associated with wear and tear. When evaluating past performance, cost accountants use tax rules that apply to broadly defined types of assets to determine allowable depreciation in their cost and profit calculations. But these depreciation allowances need not reflect the actual wear and tear on the equipment, which is likely to vary asset by asset.

Sunk Costs

Although an opportunity cost is often hidden, it should be taken into account when making economic decisions. Just the opposite is true of a **sunk cost**: an expenditure that has been made and cannot be recovered. A sunk cost is usually visible, but after it has been incurred, it should always be ignored when making future economic decisions.

Because a sunk cost cannot be recovered, it should not influence the firm's decisions. For example, consider the purchase of specialized equipment designed for a plant. Suppose the equipment can be used to do only what it was originally designed for and cannot be converted for alternative use. The expenditure on this equipment is a sunk cost. *Because it has no alternative use, its opportunity cost is zero.* Thus it should not be included as part of the firm's costs. The decision to buy this equipment may have been good or bad. It doesn't matter. It's water under the bridge and shouldn't affect current decisions.

What if, instead, the equipment could be put to other use, or could be sold or rented to another firm? In that case its use would involve an economic cost—namely, the opportunity cost of using it rather than selling or renting it to another firm.

Now consider a *prospective* sunk cost. Suppose, for example, that the firm has not yet bought the specialized equipment but is merely considering whether to do so. A prospective sunk cost is an *investment*. Here the firm must decide whether that investment in specialized equipment is *economical*—i.e., whether it will lead to a flow of revenues large enough to justify its cost. In Chapter 15, we explain in detail how to make investment decisions of this kind.

As an example, suppose a firm is considering moving its headquarters to a new city. Last year it paid \$500,000 for an option to buy a building in the city. The option gives the firm the right to buy the building at a cost of \$5,000,000, so that if it ultimately makes the purchase, its total expenditure will be \$5,500,000. Now it finds that a comparable building has become available in the same city at a price of \$5,250,000. Which building should it buy? The answer is the original building. The \$500,000 option is a cost that has been sunk and that should not affect the firm's current decision. What's at issue is spending an additional \$5,000,000 or an additional \$5,250,000. Because the economic analysis removes the sunk cost of the option from the analysis, the economic cost of the original property is \$5,000,000. The newer property, meanwhile, has an economic cost of \$5,250,000. Of course, if the new building cost \$4,750,000, the firm should buy it and forgo its option.

sunk cost Expenditure that has been made and cannot be recovered.

EXAMPLE 7.1 Choosing the Location for a New Law School Building

The Northwestern University Law School has long been located in Chicago, along the shores of Lake Michigan. However, the main campus of the university is located in the suburb of Evanston. In the mid-1970s, the law school

began planning the construction of a new building and needed to decide on an appropriate location. Should it be built on the current site, where it would remain near downtown Chicago law firms? Or should it be moved to Evanston, where it would become physically integrated with the rest of the university?

The downtown location had many prominent supporters. They argued in part that it was cost-effective to locate the new building in the city because the university already owned the land. A large parcel of land would have to be purchased in Evanston if the building were to be built there. Does this argument make economic sense?

No. It makes the common mistake of failing to appreciate opportunity costs. From an economic point of view, it is very expensive to locate downtown because the opportunity cost of the valuable lakeshore location is high: that property could have been sold for enough money to buy the Evanston land with substantial funds left over.

In the end, Northwestern decided to keep the law school in Chicago. This was a costly decision. It may have been appropriate if the Chicago location was particularly valuable to the law school, but it was inappropriate if it was made on the presumption that the downtown land was without cost.

Fixed Costs and Variable Costs

Some of the firm's costs vary with output, while others remain unchanged as long as the firm is producing any output at all. This distinction will be important when we examine the firm's profit-maximizing choice of output in the next chapter. We therefore divide **total cost (TC, or C)**—the total economic cost of production—into two components:

- **Fixed cost (FC):** A cost that does not vary with the level of output.
- **Variable cost (VC):** A cost that varies as output varies.

Depending on circumstances, fixed costs may include expenditures for plant maintenance, insurance, and perhaps a minimal number of employees. This cost remains the same no matter how much output the firm produces. Variable cost, which includes expenditures for wages, salaries, and raw materials, increases as output increases.

Fixed cost does not vary with the level of output—it must be paid even if there is no output. *The only way that a firm can eliminate its fixed costs is by going out of business.*

Which costs are variable and which are fixed depends on the time horizon that we are considering. Over a very short time horizon—say, one or two months—most costs are fixed. Over such a short period, a firm is typically obligated to receive and pay for contracted shipments of materials and cannot easily lay off workers. On the other hand, over a long time horizon—say two or three years—many costs become variable. Over a long time horizon, if the firm wants to reduce its output, it can reduce its workforce, purchase less raw material, and perhaps even sell off some of its capital.

When a firm plans a change in its operations, it usually wants to know how that change will affect its costs. Consider, for example, a problem Delta Air Lines faced recently. Delta wanted to know how its costs would change if it reduced the number of its scheduled flights by 10 percent. The answer depends on whether we are considering the short run or the long run. Over the short run—

total cost (TC or C) Total economic cost of production, consisting of fixed and variable costs.

fixed cost (FC) Cost that does not vary with the level of output.

variable cost (VC) Cost that varies as output varies.

say six months—schedules are fixed and it is difficult to lay off or discharge workers. As a result, most of Delta's short-run costs are fixed and won't be reduced significantly with the flight reduction. In the long run—say two years or more—the situation is quite different. Delta has sufficient time to sell or lease planes that are not needed and to discharge unneeded workers. In this case, most of Delta's costs are variable and thus can be reduced significantly if a 10-percent flight reduction is put in place.

Fixed versus Sunk Costs

People often confuse fixed and sunk costs. Fixed costs are costs that are paid by a firm that is in business, regardless of the level of output it produces. Such costs can include, for example, the salaries of the key executives that run the business, and expenses for their office space and support staff. Fixed costs can be avoided if the firm goes out of business—the key executives, for example, will no longer be needed. Sunk costs, on the other hand, are costs that have been incurred and *cannot be recovered*. An example is the cost of a factory with specialized equipment that is of no use in another industry. This expenditure is mostly sunk because it cannot be recovered. (Some of the cost might be recoverable if the equipment is sold for scrap.) The cost of the factory and equipment is *not* a fixed cost, because it cannot be recovered even if the firm shuts down. Suppose, on the other hand, that the firm had agreed to make payments into a worker retirement plan as long as the firm was in operation, regardless of its output or its profitability. These payments could cease only if the firm went out of business. In this case, annual payments into the retirement program should be viewed as a fixed cost.

EXAMPLE 7.2 Sunk, Fixed, and Variable Costs: Computers, Software, and Pizzas

As you progress through this book, you will see that a firm's pricing and production decisions—and its profitability—depend strongly on the structure of its costs. It is therefore important for managers to understand the characteristics of production costs and to be able to identify which costs are fixed, which are variable, and which are sunk. The relative sizes of these different cost components can vary considerably across industries. Good examples include the personal computer industry (where most costs are variable), the computer software industry (where most costs are sunk), and the pizzeria business (where most costs are fixed). Let's look at each of these in turn.

Companies like Dell, Gateway, Compaq, and IBM produce millions of personal computers every year. Because the computers they produce are very similar, competition is intense, and profitability depends critically on the ability to keep costs down. Most of these costs are variable—they increase in proportion to the number of computers produced each year. Most important is the cost of components: the microprocessor that does much of the actual computation, memory chips, hard disk drives and other storage devices, video and sound cards, etc. Typically, the majority of these components are purchased from outside suppliers in quantities that depend on the number of computers produced.

Another important part of variable cost for these companies is labor, workers are needed to assemble the computers and then to package and ship them. There is little in the way of sunk costs because the factories cost little relative to the value of the company's annual output. Likewise, there is little in the

way of fixed costs—perhaps the salaries of the top executives, some security guards, and electricity. Thus, when Dell and Gateway think about ways of reducing cost, they focus largely on getting better prices for components or reducing labor requirements—both of which are ways of reducing variable cost.

What about the software programs that run on these personal computers? Microsoft produces the Windows operating system as well as a variety of applications such as Word, Excel, and PowerPoint. But many other firms—some large and some small—also produce software programs that run on personal computers. The costs of production for such firms are quite different from those facing hardware manufacturers. In software production most costs are *sunk*. Typically, a software firm will spend a large amount of money to develop a new application program. These expenditures cannot be recovered.

Once the program is completed, the company can try to recoup its investment (and make a profit as well) by selling as many copies of the program as possible. The variable cost of producing copies of the program is very small—it is largely the cost of copying the program to floppy disks or CDs and then packaging and shipping the product. Likewise, the fixed cost of production is small. Because most costs are sunk, entering the software business can involve considerable risk. Until the development money has been spent and the product has been released for sale, an entrepreneur is unlikely to know how many copies can be sold and whether or not he will be able to make money.

Finally, let's turn to your neighborhood pizzeria. For the pizzeria, the largest component of cost is fixed. Sunk costs are fairly low because pizza ovens, chairs, tables, and dishes can be resold if the pizzeria goes out of business. Variable costs are also fairly low—mainly the ingredients for pizza (flour, tomato sauce, cheese, and pepperoni for a typical large pizza might cost \$1) and perhaps wages for a couple of workers to help produce, serve, and deliver the pizzas. Most of the cost is fixed—the opportunity cost of the owner's time (he might typically work a 60- or 70-hour week), rent, and utilities. Because of these high fixed costs, most pizzerias (which might charge \$10 for a large pizza costing about \$3 in variable cost to produce) don't make very high profits.

7.2 Cost in the Short Run

We begin our detailed analysis of cost with the short-run case. The distinction between fixed and variable costs is important here. To decide how much to produce, managers must know how variable cost increases with the level of output. It will also be helpful to consider some other measures of cost. We will use a specific example that typifies the cost situation of many firms. After we explain each of the cost concepts, we will show how they relate to the analysis in Chapter 6 of the firm's production process.

The data in Table 7.1 describe a firm with a fixed cost of \$50. Variable cost increases with output, as does total cost, which is the sum of the fixed cost in column 1 and the variable cost in column 2. From the figures given in columns 1 and 2, a number of additional cost variables can be defined.

TABLE 7.1 A Firm's Short-Run Costs

RATE OF OUTPUT (UNITS PER YEAR)	FIXED COST (DOLLARS PER YEAR)	VARIABLE COST (DOLLARS PER YEAR)	TOTAL COST (DOLLARS PER YEAR)	MARGINAL COST (DOLLARS PER UNIT)	AVERAGE FIXED COST (DOLLARS PER UNIT)	AVERAGE VARIABLE COST (DOLLARS PER UNIT)	AVERAGE TOTAL COST (DOLLARS PER UNIT)
	(FC) (1)	(VC) (2)	(TC) (3)	(MC) (4)	(AFC) (5)	(AVC) (6)	(ATC) (7)
0	50	0	50	—	—	—	—
1	50	50	100	50	50	50	100
2	50	78	128	28	25	39	64
3	50	98	148	20	16.7	32.7	49.3
4	50	112	162	14	12.5	28	40.5
5	50	130	180	18	10	26	36
6	50	150	200	20	8.3	25	33.3
7	50	175	225	25	7.1	25	32.1
8	50	204	254	29	6.3	25.5	31.8
9	50	242	292	38	5.6	26.9	32.4
10	50	300	350	58	5	30	35
11	50	385	435	85	4.5	35	39.5

Marginal Cost (MC) Marginal cost—sometimes called *incremental cost*—is the increase in cost that results from producing one extra unit of output. Because fixed cost does not change as the firm's level of output changes, marginal cost is equal to the increase in variable cost or the increase in total cost that results from an extra unit of output. We can therefore write marginal cost as

$$MC = \Delta VC / \Delta Q = \Delta TC / \Delta Q$$

Marginal cost tells us how much it will cost to expand the firm's output by one unit. In Table 7.1, marginal cost is calculated from either the variable cost (column 2) or the total cost (column 3). For example, the marginal cost of increasing output from 2 to 3 units is \$20 because the variable cost of the firm increases from \$78 to \$98. (The total cost of production also increases by \$20, from \$128 to \$148. Total cost differs from variable cost only by the fixed cost, which by definition does not change as output changes.)

Average Total Cost (ATC) Average total cost, used interchangeably with AC and with *average economic cost*, is the firm's total cost divided by its level of output, TC/Q . Thus the average total cost of producing at a rate of five units is \$36—that is, $\$180/5$. Basically, average total cost tells us the per-unit cost of production.

ATC has two components. Average fixed cost is the fixed cost (column 1 of Table 7.1) divided by the level of output, FC/Q . For example, the average fixed cost of producing 4 units of output is \$12.50 ($\$50/4$). Because fixed cost is constant, average fixed cost declines as the rate of output increases.

marginal cost (MC) Increase in cost resulting from the production of one extra unit of output.

average total cost (ATC) Firm's total cost divided by its level of output.

average fixed cost (AFC) Fixed cost divided by the level of output.

average variable cost (AVC)
Variable cost divided by the
level of output.

Average variable cost (AVC) is variable cost divided by the level of output, VC/Q . The average variable cost of producing 5 units of output is \$26—that is, $\$130/5$.

The Determinants of Short-Run Cost

Table 7.1 shows that variable and total costs increase with output. The rate at which these costs increase depends on the nature of the production process and, in particular, on the extent to which production involves diminishing returns to variable factors. Recall from Chapter 6 that diminishing returns to labor occurs when the marginal product of labor is decreasing. If labor is the only input, what happens as we increase the firm's output? To produce more output, the firm must hire more labor. Then, if the marginal product of labor decreases as the amount of labor hired is increased (owing to diminishing returns), successively greater expenditures must be made to produce output at the higher rate. As a result, variable and total costs increase as the rate of output is increased. On the other hand, if the marginal product of labor decreases only slightly as the amount of labor is increased, costs will not rise so fast when the rate of output is increased.¹

Let's look at the relationship between production and cost in more detail by concentrating on the costs of a firm that can hire as much labor as it wishes at a fixed wage w . Recall that marginal cost MC is the change in variable cost for a 1-unit change in output (i.e., $\Delta VC/\Delta Q$). But the change in variable cost is the per-unit cost of the extra labor w times the amount of extra labor needed to produce the extra output ΔL . Since $\Delta VC = w\Delta L$, it follows that

$$MC = \Delta VC/\Delta Q = w\Delta L/\Delta Q$$

Recall from Chapter 6 that the marginal product of labor MP_L is the change in output resulting from a 1-unit change in labor input, or $\Delta Q/\Delta L$. Therefore, the extra labor needed to obtain an extra unit of output is $\Delta L/\Delta Q = 1/MP_L$. As a result,

$$MC = w/MP_L \tag{7.1}$$

Equation (7.1) states that marginal cost is equal to the price of the input divided by its marginal product. Suppose, for example, that the marginal product of labor is 3 and the wage rate is \$30 per hour. In that case, 1 hour of labor will increase output by 3 units, so that 1 unit of output will require 1/3 additional hour of labor and will cost \$10. The marginal cost of producing that unit of output is \$10, which is equal to the wage, \$30, divided by the marginal product of labor, 3. A low marginal product of labor means that a large amount of additional labor is needed to produce more output, a fact that leads, in turn, to a high marginal cost. Conversely, a high marginal product means that the labor requirement is low, as is the marginal cost. More generally, whenever the marginal product of labor decreases, the marginal cost of production increases, and vice versa.²

¹ We are implicitly assuming that because labor is hired in competitive markets, the payment per unit of factor used is the same regardless of the firm's output.

² With two or more variable inputs, the relationship is more complex. The basic principle, however, still holds: The greater the productivity of factors, the less the variable cost that the firm must incur to produce any given level of output.

In §6.3, we explain that diminishing marginal returns occurs when additional inputs result in a decrease in additions to output.

The marginal product of labor is discussed in §6.3.

Diminishing Marginal Returns and Marginal Cost Diminishing marginal returns means that the marginal product of labor declines as the quantity of labor employed increases. As a result, when there are diminishing marginal returns, marginal cost will increase as output increases. This can be seen by looking at the numbers for marginal cost in Table 7.1. For output levels from 0 through 4, marginal cost is declining; for output levels from 4 through 11, however, marginal cost is increasing—a reflection of the presence of diminishing marginal returns.

The Shapes of the Cost Curves

Figure 7.1 illustrates how various cost measures change as output changes. The top part of the figure shows total cost and its two components, variable cost and

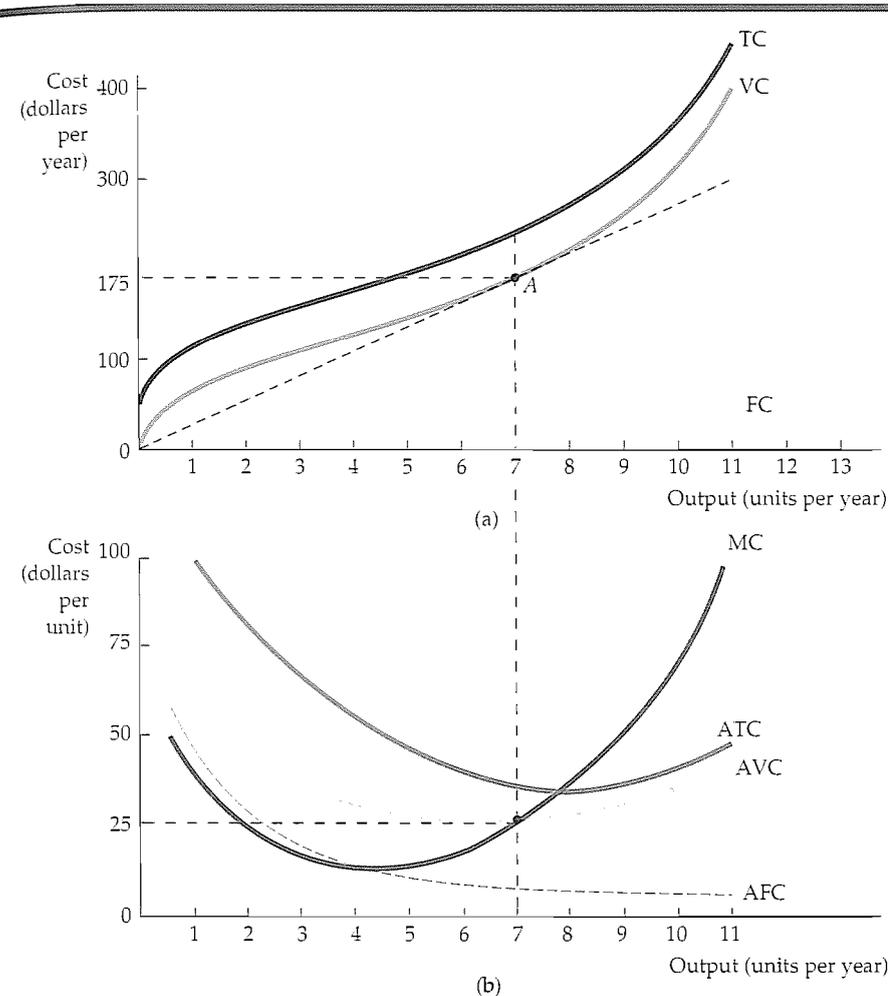


FIGURE 7.1 Cost Curves for a Firm

In (a) total cost TC is the vertical sum of fixed cost FC and variable cost VC . In (b) average total cost ATC is the sum of average variable cost AVC and average fixed cost AFC . Marginal cost MC crosses the average variable cost and average total cost curves at their minimum points.

fixed cost; the bottom part shows marginal cost and average costs. These cost curves, which are based on the information in Table 7.1, provide different kinds of information.

Observe in Figure 7.1(a) that fixed cost FC does not vary with output—it is shown as a horizontal line at \$50. Variable cost VC is zero when output is zero and then increases continuously as output increases. The total cost curve TC is determined by vertically adding the fixed cost curve to the variable cost curve. Because fixed cost is constant, the vertical distance between the two curves is always \$50.

Figure 7.1(b) shows the corresponding set of marginal and average variable cost curves.³ Because total fixed cost is \$50, the average fixed cost curve AFC falls continuously from \$50 when output is 1, toward zero for large output. The shapes of the remaining curves are determined by the relationship between the marginal and average cost curves. Whenever marginal cost lies below average cost, the average cost curve falls. Whenever marginal cost lies above average cost, the average cost curve rises. When average cost is at a minimum, marginal cost equals average cost.

Marginal and average costs are another example of the average-marginal relationship described in Chapter 6 (with respect to marginal and average product). At an output of 5 in Table 7.1, for example, the marginal cost of 18 is below the average variable cost of \$26; thus the average is lowered in response to increases in output. But when marginal cost is \$29, which is greater than average variable cost (\$25.5), the average increases as output increases. Finally, when marginal cost (\$25) and average cost (\$25) are the same, average variable cost remains unchanged (at about \$25).

The ATC curve shows the average total cost of production. Because average total cost is the sum of average variable cost and average fixed cost and the AFC curve declines everywhere, the vertical distance between the ATC and AVC curves decreases as output increases. The AVC cost curve reaches its minimum point at a lower output than the ATC curve. This follows because $MC = AVC$ at its minimum point and $MC = ATC$ at its minimum point. Because ATC is always greater than AVC and the marginal cost curve MC is rising, the minimum point of the ATC curve must lie above and to the right of the minimum point of the AVC curve.

Another way to see the relationship between the total cost curves and the average and marginal cost curves is to consider the line drawn from origin to point A in Figure 7.1(a). In that figure, the slope of the line measures average variable cost (a total cost of \$175 divided by an output of 7, or a cost per unit of \$25). Because the slope of the VC curve is the marginal cost (it measures the change in variable cost as output increases by 1 unit), the tangent to the VC curve at A is the marginal cost of production when output is 7. At A , this marginal cost of \$25 is equal to the average variable cost of \$25, because average variable cost is minimized at this output.

Note that the firm's output is measured as a flow: The firm produces a certain number of units *per year*. Thus its total cost is a flow—for example, some number of dollars per year. (Average and marginal costs, however, are measured in dollars *per unit*.) For simplicity, we will often drop the time reference, and refer to total cost in dollars and output in units. But you should remember that a firm's production of output and expenditure of cost occur over some time period. For simplicity, we will often use *cost* (C) to refer to total cost. Likewise, unless noted otherwise, we will use *average cost* (AC) to refer to average total cost.

³ The curves do not exactly match the numbers in Table 7.1. Because marginal cost represents the change in cost associated with a change in output, we have plotted the MC curve for the first unit of output by setting output equal to $\frac{1}{2}$, for the second unit by setting output equal to $1\frac{1}{2}$, and so on.

Marginal and average cost are very important concepts. As we will see in Chapter 8, they enter critically into the firm's choice of output level. Knowledge of short-run costs is particularly important for firms that operate in an environment in which demand conditions fluctuate considerably. If the firm is currently producing at a level of output at which marginal cost is sharply increasing, and if demand may increase in the future, management might want to expand production capacity to avoid higher costs.

EXAMPLE 7.3 The Short-Run Cost of Aluminum Smelting

Aluminum is a lightweight versatile metal used in a wide variety of applications, including airplanes, automobiles, packaging, and building materials. The production of aluminum begins with the mining of bauxite in such countries as Australia, Brazil, Guinea, Jamaica, and Suriname. Bauxite is an ore that contains a relatively high concentration of alumina (aluminum oxide), which is separated from the bauxite through a chemical refining process. The alumina is then converted to aluminum through a smelting process in which an electric current is used to separate the oxygen atoms from the aluminum oxide molecules. It is this smelting process—which is the most costly step in producing aluminum—that we focus on here.

All of the major aluminum producers, including Alcoa, Alcan, Reynolds, Alumax, and Kaiser, operate smelting plants. A typical smelting plant will have two production lines, each of which produces approximately 300 to 400 tons of aluminum per day. We will focus on the short-run cost of production. Thus we consider the cost of operating an existing plant because there is insufficient time in the short run to build additional plants. (It takes about four years to plan, build, and fully equip an aluminum smelting plant.)

Although the cost of a smelting plant is substantial (over \$1 billion), we will assume that the plant cannot be sold, and therefore the expenditure is sunk and can be ignored. Furthermore, because fixed costs, which are largely for administrative expenses, are relatively small, we will ignore them also. Thus we can focus entirely on short-run variable costs. Table 7.2 shows the average operating costs for a typical aluminum smelter.⁴ The cost numbers apply to a plant that runs two shifts per day to produce 600 tons of aluminum per day. If prices were sufficiently high, the firm could choose to operate the plant on a three-shifts-per-day basis by asking workers to work overtime. However, wage and maintenance costs would likely increase about 50 percent for this third shift because of the need to pay higher overtime wages. We have divided the cost components in Table 7.2 into two groups. The first group includes those costs that would remain the same at any output level, and the second includes costs that would increase if output exceeded 600 tons per day.

Note that the largest cost components for an aluminum smelter are electricity and the cost of alumina; together they represent about 60 percent of total operating costs. Because electricity, alumina, and other raw materials are used in direct proportion to the amount of aluminum produced, they represent variable costs that are constant with respect to the level of output. The costs of labor, maintenance, and freight are also proportional to the level of output,

⁴ This example is based on Kenneth S. Corts, "The Aluminum Industry in 1994," Harvard Business School Case N9-799-129, April 1999.

TABLE 7.2 Operating Costs for Aluminum Smelting (\$/ton) (based on an output of 600 tons/day)	
A Variable costs that are constant for all output levels	
Electricity	\$316
Alumina	369
Other raw materials	125
Plant power and fuel	10
Subtotal	\$820
B Variable costs that increase when output exceeds 600 tons/day	
Labor	\$150
Maintenance	120
Freight	50
Subtotal	\$320
Total operating costs	\$1140

but only when the plant operates two shifts per day. To increase output above 600 tons per day, a third shift would be necessary and would result in a 50-percent increase in the per-ton costs of labor, maintenance, and freight.

The short-run marginal cost and average variable cost curves for the smelting plant are shown in Figure 7.2. The marginal and average variable cost

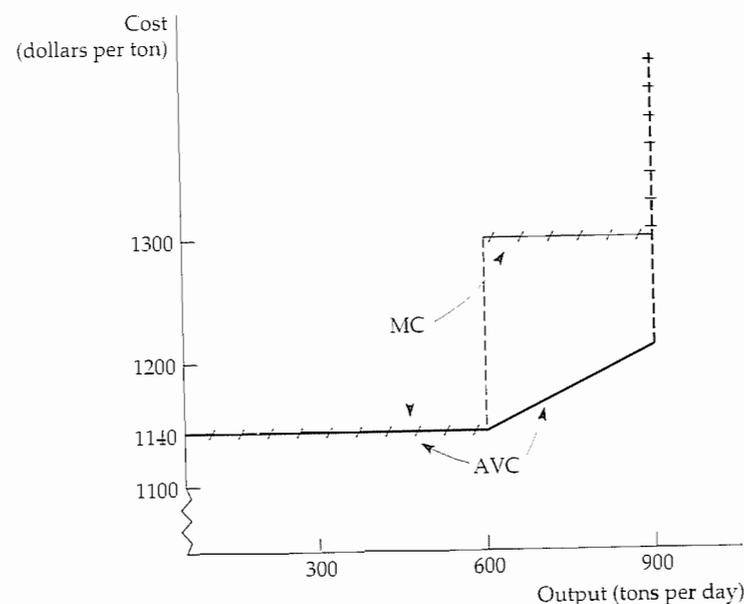


FIGURE 7.2 The Short-Run Variable Costs of Aluminum Smelting

The short-run average variable cost of smelting is constant for output levels using up to two labor shifts. When the third shift is added, marginal cost and average variable cost increase until maximum capacity is reached.

curves are horizontal at a cost of \$1140 per ton for outputs up to 600 tons per day, which represents the maximum output achievable with two shifts per day of production. As we move to the increased production of aluminum by means of a third shift, the marginal cost of labor, maintenance, and freight increases from \$320 per ton to \$480 per ton, which causes marginal cost as a whole to increase from \$1140 per ton to \$1300 per ton. As the figure shows, the increases in marginal costs cause average costs to increase as well. Finally, when output reaches 900 tons per day, an absolute capacity constraint is reached, at which point the marginal and average costs of production become infinite.

7.3 Cost in the Long Run

In the long run, the firm can change all its inputs. In this section we show how a firm chooses the combination of inputs that minimizes the cost of producing a given output. We will also examine the relationship between long-run cost and the level of output. We begin by taking a careful look at the firm's cost of using capital equipment. We then show how this cost, along with the cost of labor, enters into the production decision.

The User Cost of Capital

Firms often rent or lease equipment, buildings, and other capital used in the production process. On other occasions, the capital is purchased. In our analysis, however, it will be useful to treat capital as though it were rented, even if it was, in fact, purchased. An illustration will help to explain how and why we do this. Following on our previous example, let's suppose that Delta Airlines is thinking about purchasing a new Boeing 777 airplane for \$150 million. Even though Delta would pay a large sum for the airplane now, for economic purposes the purchase price can be allocated or *amortized* across the life of the airplane. This will allow Delta to compare its revenues and costs on an *annual flow basis*. We will assume that the life of the airplane is 30 years; the amortized cost is therefore \$5 million per year. The \$5 million can be viewed as the *annual economic depreciation* for the airplane.

So far, we have ignored the fact that had the firm not purchased the airplane, it could have earned interest on its \$150 million. This forgone interest is an *opportunity cost* that must be accounted for. Therefore, the **user cost of capital**—the annual cost of owning and using the airplane instead of selling it or never buying it in the first place—is given by the *sum of the economic depreciation and the interest (i.e., the financial return) that could have been earned had the money been invested elsewhere.*⁵ Formally,

$$\text{User Cost of Capital} = \text{Economic Depreciation} + (\text{Interest Rate})(\text{Value of Capital})$$

user cost of capital Sum of the annual cost of owning and using a capital asset, equal to economic depreciation plus forgone interest.

⁵ More precisely, the financial return should reflect an investment with similar risk. The interest rate, therefore, should include a risk premium. We discuss this point in Chapter 15.

In our example, economic depreciation on the airplane is \$5 million per year. Suppose Delta could have earned a return of 10 percent had it invested its money elsewhere. In that case, the user cost of capital is \$5 million + (.10)(\$150 million - depreciation). As the plane depreciates over time, its value declines, as does the opportunity cost of the financial capital that is invested in it. For example, at the time of purchase, looking forward for the first year, the user cost of capital is \$5 million + (.10)(\$150 million) = \$20 million. In the tenth year of ownership, the airplane, which will have depreciated by \$50 million, will be worth \$100 million. At that point, the user cost of capital will be \$5 million + (.10)(\$100 million) = \$15 million per year.

We can also express the user cost of capital as a *rate* per dollar of capital:

$$r = \text{Depreciation rate} + \text{Interest rate}$$

For our airplane example, the depreciation rate is $1/30 = 3.33$ percent per year. If Delta could have earned a rate of return of 10 percent per year, its user cost of capital would be $r = 3.33 + 10 = 13.33$ percent per year.

In the long run, the firm can change all its inputs. We will now show how the firm chooses the combination of inputs that minimizes the cost of producing a given output. We will then examine the relationship between long-run cost and the level of output.

The Cost-Minimizing Input Choice

We now turn to a fundamental problem that all firms face: *how to select inputs to produce a given output at minimum cost*. For simplicity, we will work with two variable inputs: labor (measured in hours of work per year) and capital (measured in hours of use of machinery per year).

The amount of labor and capital that the firm uses will depend, of course, on the prices of these inputs. We will assume that there are competitive markets for both inputs, so that their prices are unaffected by what the firm does. (In Chapter 14 we will examine labor markets that are not competitive.) In this case, the price of labor is simply the *wage rate*, w . But what about the price of capital?

The Price of Capital In the long run, the firm can adjust the amount of capital it uses. (Even if the capital includes specialized machinery that has no alternative use, expenditures on this machinery are not yet sunk and must be taken into account; the firm is deciding *prospectively* how much capital to obtain. Unlike labor expenditures, however, large initial expenditures on capital are necessary. In order to compare the firm's expenditure on capital with its ongoing cost of labor, we want to express this capital expenditure as a *flow*—e.g., in dollars per year. To do this, we must amortize the expenditure by spreading it over the lifetime of the capital, and we must also account for the forgone interest that the firm could have earned by investing the money elsewhere. As we have just seen, this is exactly what we do when we calculate the *user cost of capital*. As above, the price of capital is its *user cost*, given by $r = \text{Depreciation rate} + \text{Interest rate}$.

Rental Rate of Capital Sometimes capital is rented rather than purchased. An example is office space in a large office building. In this case, the price of capital is its **rental rate**—i.e., the cost per year for renting a unit of capital.

rental rate Cost per year of renting one unit of capital.

Does this mean that we must distinguish between capital that is rented and capital that is purchased when we determine the price of capital? No. If the capital market is competitive (as we have assumed it is), *the rental rate should be equal to the user cost, r* . Why? Because in a competitive market, firms that own capital (e.g., the owner of the large office building) expect to earn a competitive return when they rent it—namely, the rate of return that they could have earned by investing their money elsewhere, plus an amount to compensate for the depreciation of the capital. *This competitive return is the user cost of capital.*

Many textbooks simply assume that all capital is rented at a rental rate r . As we have just seen, this assumption is reasonable. However, you should now understand *why* it is reasonable: *Capital that is purchased can be treated as though it were rented at a rental rate equal to the user cost of capital.*

For the remainder of this chapter, we will therefore assume that the firm rents all of its capital at a rental rate, or “price,” r , just as it hires labor at a wage rate, or “price,” w . We can now focus on how a firm takes these prices into account when determining how much capital and labor to utilize.⁶

The Isocost Line

We begin by looking at the cost of hiring factor inputs, which can be represented by a firm's isocost lines. An **isocost line** shows all possible combinations of labor and capital that can be purchased for a given total cost. To see what an isocost line looks like, recall that the total cost C of producing any particular output is given by the sum of the firm's labor cost wL and its capital cost rK :

$$C = wL + rK \quad (7.2)$$

For each different level of total cost, equation (7.2) describes a different isocost line. In Figure 7.3, for example, the isocost line C_0 describes all possible combinations of labor and capital that cost a total of C_0 to hire.

If we rewrite the total cost equation as an equation for a straight line, we get

$$K = C/r - (w/r)L$$

It follows that the isocost line has a slope of $\Delta K/\Delta L = -(w/r)$, which is the ratio of the wage rate to the rental cost of capital. Note that this slope is similar to the slope of the budget line that the consumer faces (because it is determined solely by the prices of the goods in question, whether inputs or outputs). It tells us that if the firm gave up a unit of labor (and recovered w dollars in cost) to buy w/r units of capital at a cost of r dollars per unit, its total cost of production would remain the same. For example, if the wage rate were \$10 and the rental cost of capital \$5, the firm could replace one unit of labor with two units of capital with no change in total cost.

⁶ It is possible, of course, that input prices might increase with demand because of overtime or a relative shortage of capital equipment. We discuss the possibility of a relationship between the price of factor inputs and the quantities demanded by a firm in Chapter 14.

isocost line Graph showing all possible combinations of labor and capital that can be purchased for a given total cost.

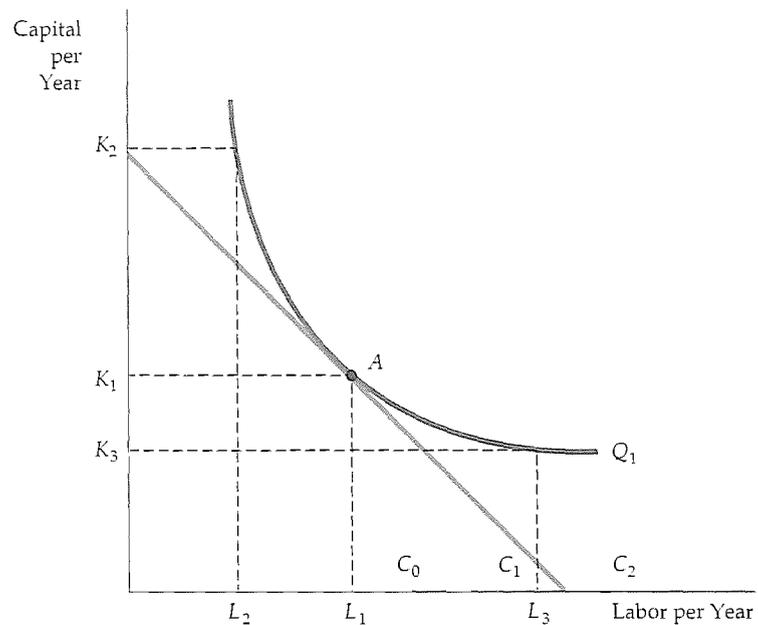


FIGURE 7.3 Producing a Given Output at Minimum Cost

Isocost curves describe the combination of inputs to production that cost the same amount to the firm. Isocost curve C_1 is tangent to isoquant Q_1 at A and shows that output Q_1 can be produced at minimum cost with labor input L_1 and capital input K_1 . Other input combinations— L_2, K_2 and L_3, K_3 —yield the same output at higher cost.

Choosing Inputs

Suppose we wish to produce at an output level Q_1 . How can we do so at minimum cost? Look at the firm's production isoquant, labeled Q_1 , in Figure 7.3. The problem is to choose the point on this isoquant that minimizes total cost.

Figure 7.3 illustrates the solution to this problem. Suppose the firm were to spend C_0 on inputs. Unfortunately, no combination of inputs can be purchased for expenditure C_0 that will allow the firm to achieve output Q_1 . However, output Q_1 can be achieved with the expenditure of C_2 , either by using K_2 units of capital and L_2 units of labor or by using K_3 units of capital and L_3 units of labor. But C_2 is not the minimum cost. The same output Q_1 can be produced more cheaply, at a cost of C_1 , by using K_1 units of capital and L_1 units of labor. In fact, isocost line C_1 is the lowest isocost line that allows output Q_1 to be produced. The point of tangency of the isoquant Q_1 and the isocost line C_1 at point A tells us the cost-minimizing choice of inputs, L_1 and K_1 , which can be read directly from the diagram. At this point, the slopes of the isoquant and the isocost line are just equal.

When the expenditure on all inputs increases, the slope of the isocost line does not change—because the prices of the inputs have not changed. The intercept, however, increases. Suppose that the price of one of the inputs, such as labor, were to increase. In that case, the slope of the isocost line $-(w/r)$ would increase in magnitude and the isocost line would become steeper.

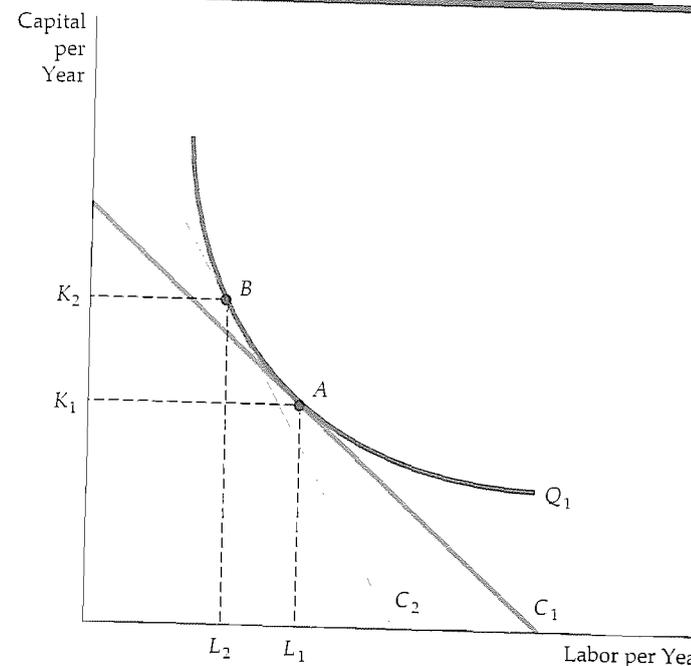


FIGURE 7.4 Input Substitution When an Input Price Changes

Facing an isocost curve C_1 , the firm produces output Q_1 at point A using L_1 units of labor and K_1 units of capital. When the price of labor increases, the isocost curves become steeper. Output Q_1 is now produced at point B on isocost curve C_2 by using L_2 units of labor and K_2 units of capital.

Figure 7.4 shows this. Initially, the isocost line is C_1 , and the firm minimizes its costs of producing output Q_1 at A by using L_1 units of labor and K_1 units of capital. When the price of labor increases, the isocost line becomes steeper. The isocost line C_2 reflects the higher price of labor. Facing this higher price of labor, the firm minimizes its cost of producing output Q_1 by producing at B , using L_2 units of labor and K_2 units of capital. The firm has responded to the higher price of labor by substituting capital for labor in the production process.

How does the isocost line relate to the firm's production process? Recall that in our analysis of production technology, we showed that the marginal rate of technical substitution MRTS of labor for capital is the negative of the slope of the isoquant and is equal to the ratio of the marginal products of labor and capital:

$$\text{MRTS} = -\Delta K/\Delta L = \text{MP}_L/\text{MP}_K \quad (7.3)$$

Above, we noted that the isocost line has a slope of $\Delta K/\Delta L = -w/r$. It follows that when a firm minimizes the cost of producing a particular output, the following condition holds:

$$\text{MP}_L/\text{MP}_K = w/r$$

In §6.2, we explain that the MRTS is the amount by which the input of capital can be reduced when one extra unit of labor is used, so that output remains constant.

We can rewrite this condition slightly as follows:

$$MP_L/w = MP_K/r \quad (7.4)$$

MP_L/w is the additional output that results from spending an additional dollar for labor. Suppose that the wage rate is \$10 and that adding a worker to the production process will increase output by 20 units. The additional output per dollar spent on an additional worker will be $20/10 = 2$ units of output per dollar. Similarly, MP_K/r is the additional output that results from spending an additional dollar for capital. Therefore, equation (7.4) tells us that a cost-minimizing firm should choose its quantities of inputs so that the last dollar's worth of any input added to the production process yields the same amount of extra output.

Why must this condition hold for cost minimization? Suppose that in addition to the \$10 wage rate, the rental rate on capital is \$2. Suppose also that adding a unit of capital will increase output by 20 units. In that case, the additional output per dollar of capital input would be $20/\$2 = 10$ units of output per dollar. Because a dollar spent for capital is five times more productive than a dollar spent for labor, the firm will want to use more capital and less labor. If the firm reduces labor and increases capital, its marginal product of labor will rise and its marginal product of capital will fall. Eventually, the point will be reached where the production of an additional unit of output costs the same regardless of which additional input is used. At that point the firm is minimizing its cost.

EXAMPLE 7.4 The Effect of Effluent Fees on Input Choices

Steel plants are often built on or near rivers. A river offers readily available, inexpensive transportation for both the iron ore that goes into the production process and the finished steel itself. Unfortunately, it also provides a cheap disposal method for by-products of the production process, called *effluent*. For example, a steel plant processes its iron ore for use in blast furnaces by grinding taconite deposits into a fine consistency. During this process, the ore is extracted by a magnetic field as a flow of water and fine ore passes through the plant. One by-product of this process—fine taconite particles—can be dumped in the river at relatively little cost to the firm. Alternative removal methods or private treatment plants are relatively expensive.

Because the taconite particles are a nondegradable waste that can harm vegetation and fish, the Environmental Protection Agency (EPA) has imposed an effluent fee—a per-unit fee that the steel firm must pay for the effluent that goes into the river. How should the manager of a steel plant deal with the imposition of this effluent fee to minimize the costs of production?

Suppose that without regulation the plant is producing 2000 tons of steel per month, using 2000 machine-hours of capital and 10,000 gallons of water (which contains taconite particles when returned to the river). The manager estimates that a machine-hour costs \$40 and dumping each gallon of wastewater in the river costs \$10. The total cost of production is therefore \$180,000: \$80,000 for capital and \$100,000 for wastewater. How should the manager respond to an EPA-imposed effluent fee of \$10 per gallon of wastewater dumped? The manager

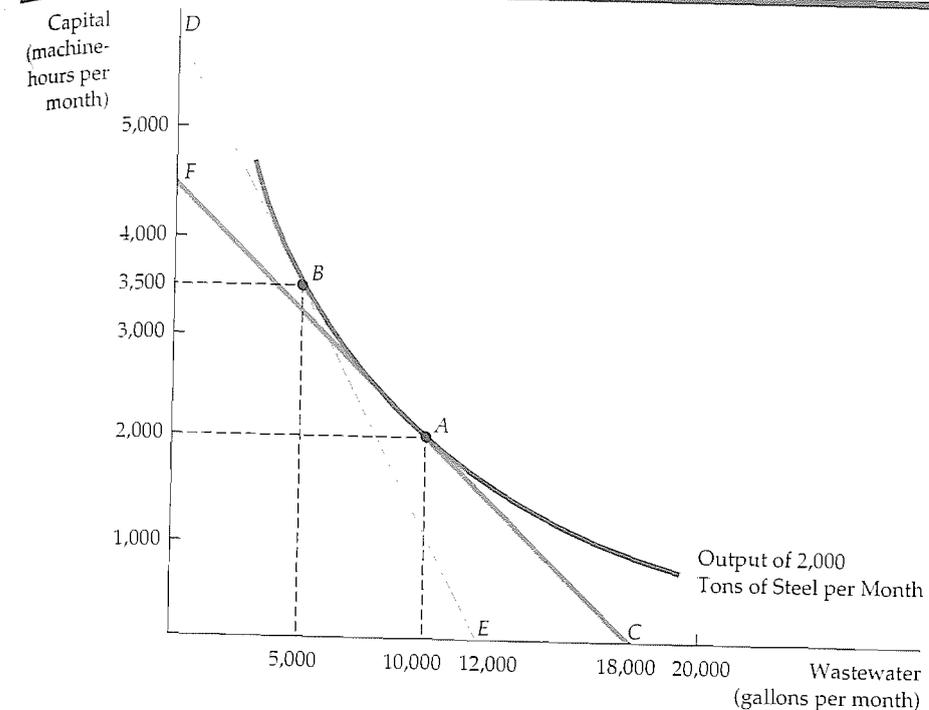


FIGURE 7.5 The Cost-Minimizing Response to an Effluent Fee

When the firm is not charged for dumping its wastewater in a river, it chooses to produce a given output using 10,000 gallons of wastewater and 2000 machine-hours of capital at *A*. However, an effluent fee raises the cost of wastewater, shifts the isocost curve from *FC* to *DE*, and causes the firm to produce at *B*—a process that results in much less effluent.

knows that there is some flexibility in the production process. If the firm puts into place more expensive effluent treatment equipment, the firm can achieve the same output with less wastewater.

Figure 7.5 shows the cost-minimizing response. The vertical axis measures the firm's input of capital in machine-hours per month—the horizontal axis measures the quantity of wastewater in gallons per month. First, consider the level at which the firm produces when there is no effluent fee. Point *A* represents the input of capital and the level of wastewater that allows the firm to produce its quota of steel at minimum cost. Because the firm is minimizing cost, *A* lies on the isocost line *FC*, which is tangent to the isoquant. The slope of the isocost line is equal to $-\$10/\$40 = -0.25$ because a unit of capital costs four times more than a unit of wastewater.

When the effluent fee is imposed, the cost of wastewater increases from \$10 per gallon to \$20: For every gallon of wastewater (which costs \$10), the firm has to pay the government an additional \$10. The effluent fee therefore increases the cost of wastewater relative to capital. To produce the same output at the lowest possible cost, the manager must choose the isocost line with a slope of $-\$20/\$40 = -0.5$ that is tangent to the isoquant. In Figure 7.5, *DE* is the appropriate isocost line, and *B* gives the appropriate choice of capital and wastewater. The move from *A* to *B* shows that with an effluent fee the use of an alternative production technology that emphasizes the use of capital (3500

machine-hours) and uses less wastewater (5000 gallons) is cheaper than the original process which did not emphasize recycling. Note that the total cost of production has increased to \$240,000: \$140,000 for capital, \$50,000 for wastewater, and \$50,000 for the effluent fee.)

We can learn two lessons from this decision. First, the more easily factors can be substituted in the production process—that is, the more easily the firm can deal with its taconite particles without using the river for waste treatment—the more effective the fee will be in reducing effluent. Second, the greater the degree of substitution, the less the firm will have to pay. In our example, the fee would have been \$100,000 had the firm not changed its inputs. However, the steel company pays only a \$50,000 fee by moving production from A to B.

Cost Minimization with Varying Output Levels

In the previous section we saw how a cost-minimizing firm selects a combination of inputs to produce a given level of output. Now we extend this analysis to see how the firm's costs depend on its output level. To do this we determine the firm's cost-minimizing input quantities for each output level and then calculate the resulting cost.

The cost-minimization exercise yields the result illustrated by Figure 7.6. We have assumed that the firm can hire labor L at $w = \$10/\text{hour}$ and rent a unit of capital K for $r = \$20/\text{hour}$. Given these input costs, we have drawn three of the firm's isocost lines. Each isocost line is given by the following equation:

$$C = (\$10/\text{hour})(L) + (\$20/\text{hour})(K)$$

In the figure, the lowest (unlabeled) line represents a cost of \$1,000; the middle line \$2,000, and the highest line \$3,000.

You can see that each of the points A, B, and C in Figure 7.6(a) is a point of tangency between an isocost curve and an isoquant. Point B, for example, shows us that the lowest-cost way to produce 200 units of output is to use 100 units of labor and 50 units of capital; this combination lies on the \$2,000 isocost line. Similarly, the lowest-cost way to produce 100 units of output (the lowest unlabeled isoquant) is \$1,000 (at point A, $L = 50$, $K = 25$); the least-cost means of getting 300 units of output is \$3,000 (at point C, $L = 150$, $K = 75$).

The curve passing through the points of tangency between the firm's isocost lines and its isoquants is its *expansion path*. The *expansion path* describes the combinations of labor and capital that the firm will choose to minimize costs at each output level. As long as the use of both labor and capital increases with output, the curve will be upward sloping. In this particular case we can easily calculate the slope of the line. As output increases from 100 to 200 units, capital increases from 25 to 50 units, while labor increases from 50 to 100 units. For each level of output, the firm uses half as much capital as labor. Therefore, the expansion path is a straight line with a slope equal to

$$\Delta K/\Delta L = (50 - 25)/(100 - 50) = \frac{1}{2}$$

The Expansion Path and Long-Run Costs

The firm's expansion path contains the same information as its long-run total cost curve, $C(q)$. This can be seen in Figure 7.6(b). To move from the expansion path to the cost curve, we follow three steps:

expansion path Curve passing through points of tangency between a firm's isocost lines and its isoquants.

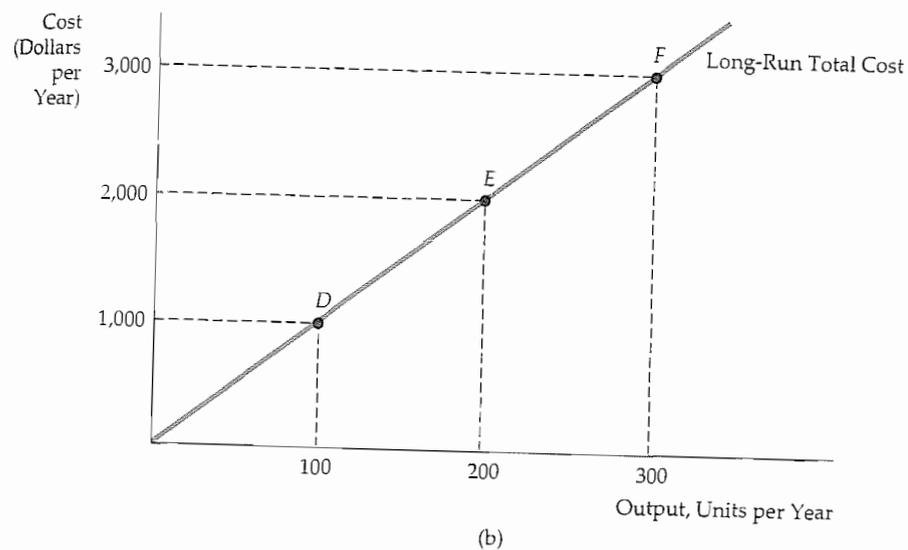
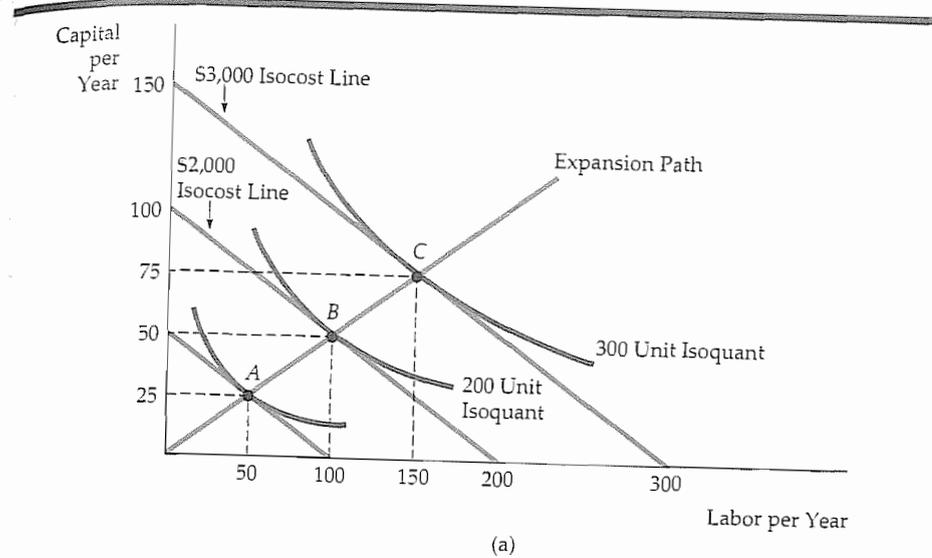


FIGURE 7.6 A Firm's Expansion Path and Long-Run Total Cost Curve

In (a), the expansion path (from the origin through points A, B, and C) illustrates the least-cost combinations of labor and capital that can be used to produce each level of output in the long run—i.e., when both inputs to production can be varied. In (b), the corresponding long-run total cost curve (from the origin through points D, E, and F) measures the least cost of producing the three output levels shown in (a).

1. Choose an output level represented by an isoquant in Figure 7.6(a). Then find the point of tangency of that isoquant with an isocost line.
2. From the chosen isocost line determine the minimum cost of producing the output level that has been selected.
3. Graph the output-cost combination in Figure 7.6(b).

Suppose we begin with an output of 100 units. The point of tangency of the 100-unit isoquant with an isocost line is given by point *A* in Figure 7.6(a). Because *A* lies on the \$1,000 isocost line, we know that the minimum cost of producing an output of 100 units in the long-run is \$1,000. We graph this combination of 100 units of output and \$1,000 cost as point *D* in Figure 7.6(b). Point *D* thus represents the \$1,000 cost of producing 100 units of output. Similarly, point *E* represents the \$2,000 cost of producing 200 units, which corresponds to point *B* on the expansion path. Finally, point *F* represents the \$3,000 cost of 300 units corresponding to point *C*. Repeating these steps for every level of output gives the *long-run total cost curve* in Figure 7.6(b)—i.e., the minimum long-run cost of producing each level of output.

In this particular example, the long-run total cost curve is a straight line. Why? Because there are constant returns to scale in production: As inputs increase proportionately, so do outputs. As we will see in the next section, the shape of the expansion path provides information about how costs change with the scale of the firm's operation.

7.4 Long-Run versus Short-Run Cost Curves

We saw earlier (see Figure 7.1) that short-run average cost curves are U-shaped. We will see that long-run average cost curves can also be U-shaped, but different economic factors explain the shapes of these curves. In this section, we discuss long-run average and marginal cost curves and highlight the differences between the curves and their short-run counterparts.

The Inflexibility of Short-Run Production

Recall that we defined the long run as occurring when all inputs to the firm are variable. In the long run, the firm's planning horizon is long enough to allow for a change in plant size. This added flexibility allows the firm to produce at a lower average cost than in the short run. To see why, we might compare the situation in which capital and labor are both flexible to the case in which capital is fixed in the short run.

Figure 7.7 shows the firm's production isoquants. The firm's *long-run expansion path* is the straight line from the origin that corresponds to the expansion path in Figure 7.6. Now, suppose capital is fixed at a level K_1 in the short run. To produce output Q_1 , the firm would minimize costs by choosing labor equal to L_1 , corresponding to the point of tangency with the isocost line *AB*. The inflexibility appears when the firm decides to increase its output to Q_2 without increasing its use of capital. If capital were not fixed, it would produce this output with capital K_2 and labor L_2 . Its cost of production would be reflected by isocost line *CD*.

However, the fact that capital is fixed forces the firm to increase its output by using capital K_1 and labor L_3 at point *P*. Point *P* lies on the isocost line *EF*, which represents a higher cost than isocost line *CD*. Why is the cost of production higher when capital is fixed? Because the firm is unable to substitute relatively inexpensive capital for more costly labor when it expands production. This inflexibility is reflected in the *short-run expansion path*, which begins as a line

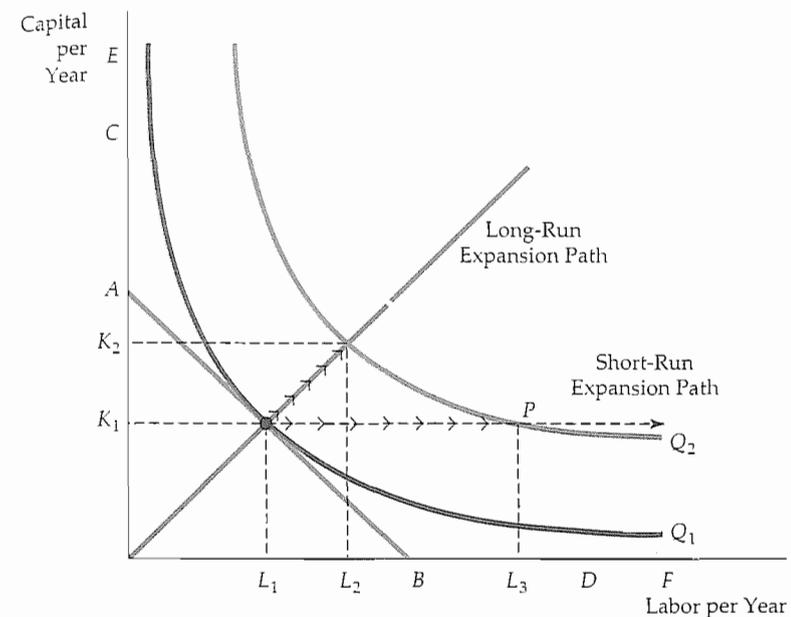


FIGURE 7.7 The Inflexibility of Short-Run Production

When a firm operates in the short run, its cost of production may not be minimized because of inflexibility in the use of capital inputs. Output is initially at level Q_1 . In the short run, output Q_2 can be produced only by increasing labor from L_1 to L_3 because capital is fixed at K_1 . In the long run, the same output can be produced more cheaply by increasing labor from L_1 to L_2 and capital from K_1 to K_2 .

from the origin and then becomes a horizontal line when the capital input reaches K_1 .

Long-Run Average Cost

In the long run, the ability to change the amount of capital allows the firm to reduce costs. To see how costs vary as the firm moves along its expansion path in the long run, we can look at the long-run average and marginal cost curves.⁷ The most important determinant of the shape of the long-run average and marginal cost curves is the relationship between the scale of the firm's operation and the inputs that are required to minimize the firm's costs. Suppose, for example, that the firm's production process exhibits constant returns to scale at all input levels. In this case, a doubling of inputs leads to a doubling of output. Because input prices remain unchanged as output increases, the average cost of production must be the same for all levels of output.

Suppose instead that the firm's production process is subject to increasing returns to scale: A doubling of inputs leads to more than a doubling of output. In that case, the average cost of production falls with output because a doubling of

⁷ We saw that in the short run, the shapes of the average and marginal cost curves were determined primarily by diminishing returns. As we showed in Chapter 6, diminishing returns to each factor is consistent with constant (or even increasing) returns to scale.

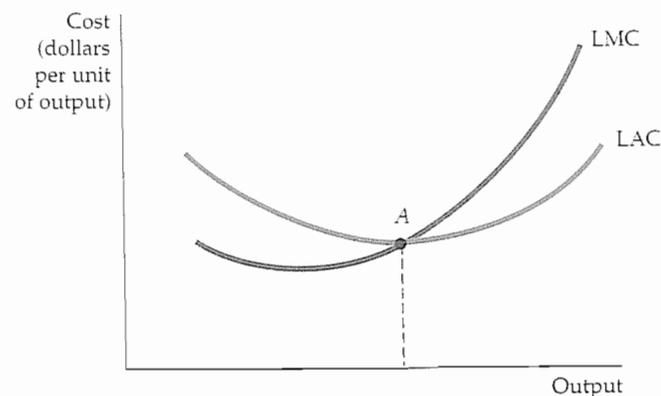


FIGURE 7.8 Long-Run Average and Marginal Cost

When a firm is producing at an output at which the long-run average cost LAC is falling, the long-run marginal cost LMC is less than LAC. Conversely, when LAC is increasing, LMC is greater than LAC. The two curves intersect at A, where the LAC curve achieves its minimum.

costs is associated with a more than twofold increase in output. By the same logic, when there are decreasing returns to scale, the average cost of production must be increasing with output.

We saw that the long-run total cost curve associated with the expansion path in Figure 7.6(a) was a straight line from the origin. In this constant-returns-to-scale case, the long-run average cost of production is constant: It is unchanged as output increases. For an output of 100, long-run average cost is $\$1,000/100 = \10 per unit. For an output of 200, long-run average cost is $\$2,000/200 = \10 per unit; for an output of 300, average cost is also $\$10$ per unit. Because a constant average cost means a constant marginal cost, the long-run average and marginal cost curves are given by a horizontal line at a $\$10$ /unit cost.

Recall that in the last chapter we examined a firm's production technology that exhibits first increasing returns to scale, then constant returns to scale, and eventually decreasing returns to scale. Figure 7.8 shows a typical **long-run average cost curve (LAC)** consistent with this description of the production process. Like the **short-run average cost curve**, the long-run average cost curve is U-shaped, but the source of the U-shape is increasing and decreasing returns to scale, rather than diminishing returns to a factor of production.

The **long-run marginal cost curve (LMC)** can be determined from the long-run average cost curve; it measures the change in long-run total costs as output is increased incrementally. LMC lies below the long-run average cost curve when LAC is falling and above it when LAC is rising.⁵ The two curves intersect at A, where the long-run average cost curve achieves its minimum. In the special case in which LAC is constant, LAC and LMC are equal.

⁵ Recall that $AC = TC/Q$. It follows that, $\Delta AC/\Delta Q = [Q(\Delta TC/\Delta Q) - TC]/Q^2 = (MC - AC)/Q$. Clearly, when AC is increasing, $\Delta AC/Q$ is positive and $MC > AC$. Correspondingly, when AC is decreasing, $\Delta AC/\Delta Q$ is negative and $MC < AC$.

long-run average cost curve (LAC) Curve relating average cost of production to output when all inputs, including capital, are variable.

short-run average cost curve (SAC) Curve relating average cost of production to output when level of capital is fixed.

long-run marginal cost curve (LMC) Change in long-run total cost as output is increased incrementally by 1 unit.

Economies and Diseconomies of Scale

In the long run, it may be in the firm's interest to change the input proportions as the level of output changes. When input proportions do change, the firm's expansion path is no longer a straight line, and the concept of returns to scale no longer applies. Rather, we say that a firm enjoys **economies of scale** when it can double its output for less than twice the cost. Correspondingly, there are **diseconomies of scale** when a doubling of output requires more than twice the cost. The term *economies of scale* includes increasing returns to scale as a special case, but it is more general because it reflects input proportions that change as the firm changes its level of production. In this more general setting, a U-shaped long-run average cost curve characterizes the firm facing economies of scale for relatively low output levels and diseconomies of scale for higher levels.

Economies of scale are often measured in terms of a cost-output elasticity, E_C . E_C is the percentage change in the cost of production resulting from a 1-percent increase in output:

$$E_C = (\Delta C/C)/(\Delta Q/Q) \quad (7.5)$$

To see how E_C relates to our traditional measures of cost, rewrite equation (7.5) as follows:

$$E_C = (\Delta C/\Delta Q)/(C/Q) = MC/AC \quad (7.6)$$

Clearly, E_C is equal to 1 when marginal and average costs are equal. In that case, costs increase proportionately with output, and there are neither economies nor diseconomies of scale (constant returns to scale would apply if input proportions were fixed). When there are economies of scale (when costs increase less than proportionately with output), marginal cost is less than average cost (both are declining) and E_C is less than 1. Finally, when there are diseconomies of scale, marginal cost is greater than average cost and E_C is greater than 1.

The Relationship Between Short-Run and Long-Run Cost

Figures 7.9 and 7.10 show the relationship between short-run and long-run cost. Assume that a firm is uncertain about the future demand for its product and is considering three alternative plant sizes. The short-run average cost curves for the three plants are given by SAC_1 , SAC_2 , and SAC_3 in Figure 7.9. The decision is important because, once built, the firm may not be able to change the plant size for some time.

Figure 7.9 shows the case in which there are constant returns to scale in the long run. If the firm expects to produce Q_1 units of output, then it should build the smallest plant. Its average cost of production would be $\$10$; this is the minimum cost because the short-run marginal cost SMC crosses short-run average cost SAC when both equal $\$10$. If the firm expects to produce Q_2 , the middle-sized plant is best, and its average cost of production is again $\$10$. If it is to produce Q_3 , the third plant is best. With only these plant sizes, any production choice between Q_1 and Q_2 will entail an increase in the average cost of production, as will any level of production between Q_2 and Q_3 .

What is the firm's long-run cost curve? In the long run, the firm can change the size of its plant. Thus if it was initially producing Q_1 and wanted to increase output to Q_2 or Q_3 , it could do so with no increase in average cost. With three

economies of scale Output can be doubled for less than a doubling of cost.

In §6.4, we explain that increasing returns to scale occurs when output more than doubles when inputs are doubled proportionately.

diseconomies of scale A doubling of output requires more than a doubling of cost.

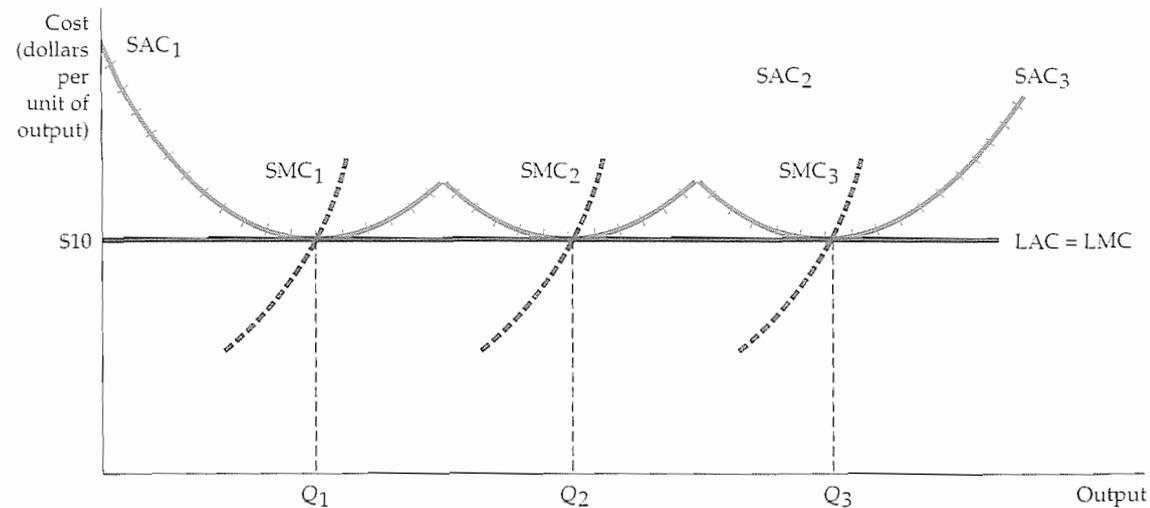


FIGURE 7.9 Long-Run Cost with Constant Returns to Scale

The long-run average cost curve LAC, which is identical to the long-run marginal cost curve LMC, is the envelope of the short-run average cost curves (SAC_1 , SAC_2 , and SAC_3 are shown). With constant returns to scale, the long-run average cost curve consists of the minimum points of the short-run average cost curves.

possible plant sizes, the long-run average cost curve is therefore given by the crosshatched portions of the short-run average cost curves because these show the minimum cost of production for any output level. The long-run average cost curve is the *envelope* of the short-run average cost curves—it envelops or surrounds the short-run curves.

Now suppose that there are many choices of plant size, each having a short-run average cost curve with a minimum of \$10. Again, the long-run average cost curve is the envelope of the short-run curves. In Figure 7.9 it is the straight line LAC. Whatever the firm wants to produce, it can choose the plant size (and the mix of capital and labor) that allows it to produce that output at the minimum average cost of \$10.

With economies or diseconomies of scale, the analysis is essentially the same, but the long-run average cost curve is no longer a horizontal line. Figure 7.10 illustrates the typical case in which three plant sizes are possible; the minimum average cost is lowest for a medium-sized plant. The long-run average cost curve, therefore, exhibits economies of scale initially but exhibits diseconomies at higher output levels. Once again, the crosshatched lines show the long-run average cost associated with the three plants.

To clarify the relationship between the short-run and the long-run cost curves, consider a firm that wants to produce output Q_1 in Figure 7.10. If it builds a small plant, the short-run average cost curve SAC_1 is relevant. The average cost of production (at B on SAC_1) is \$8. A small plant is a better choice than a medium-sized plant with an average cost of production of \$10 (A on curve SAC_2). Point B would therefore become one point on the long-run cost function when only three plant sizes are possible. If plants of other sizes could be built, and if at least one size allowed the firm to produce Q_1 at less than \$8 per unit, then B would no longer be on the long-run cost curve.

In Figure 7.10, the envelope that would arise if plants of any size could be built is given by the LAC curve, which is U-shaped. Note, once again, that the

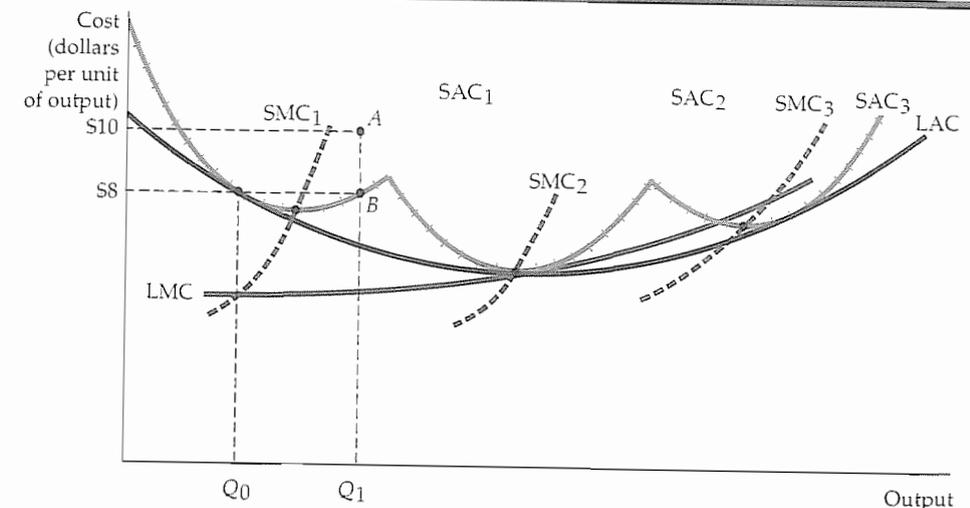


FIGURE 7.10 Long-Run Cost with Economies and Diseconomies of Scale

The long-run average cost curve LAC is the envelope of the short-run average cost curves SAC_1 , SAC_2 , and SAC_3 . With economies and diseconomies of scale, the minimum points of the short-run average cost curves do not lie on the long-run average cost curve.

LAC curve never lies above any of the short-run average cost curves. Also note that because there are economies and diseconomies of scale in the long run, the points of minimum average cost of the smallest and largest plants do *not* lie on the long-run average cost curve. For example, a small plant operating at minimum average cost is not efficient because a larger plant can take advantage of increasing returns to scale to produce at a lower average cost.

Finally, note that the long-run marginal cost curve LMC is not the envelope of the short-run marginal cost curves. Short-run marginal costs apply to a particular plant; long-run marginal costs apply to all possible plant sizes. Each point on the long-run marginal cost curve is the short-run marginal cost associated with the most cost-efficient plant. Consistent with this relationship, SMC_1 intersects LMC in Figure 7.10 at the output level Q_0 at which SAC_1 is tangent to LAC.

7.5 Production with Two Outputs—Economies of Scope

Many firms produce more than one product. Sometimes a firm's products are closely linked to one another: A chicken farm, for instance, produces poultry and eggs, an automobile company produces automobiles and trucks, and a university produces teaching and research. At other times, firms produce physically unrelated products. In both cases, however, a firm is likely to enjoy production or cost advantages when it produces two or more products. These advantages could result from the joint use of inputs or production facilities, joint marketing programs, or possibly the cost savings of a common administration. In some

cases, the production of one product gives an automatic and unavoidable by-product that is valuable to the firm. For example, sheet metal manufacturers produce scrap metal and shavings they can sell.

Product Transformation Curves

To study the economic advantages of joint production, let's consider an automobile company that produces two products, cars and tractors. Both products use capital (factories and machinery) and labor as inputs. Cars and tractors are not typically produced at the same plant, but they do share management resources, and both rely on similar machinery and skilled labor. The managers of the company must choose how much of each product to produce. Figure 7.11 shows two product transformation curves, each showing the various combinations of cars and tractors that can be produced with a given input of labor and machinery. Curve O_1 describes all combinations of the two outputs that can be produced with a relatively low level of inputs and Curve O_2 describes the output combinations associated with twice the inputs.

The product transformation curve has a negative slope because in order to get more of one output, the firm must give up some of the other output. For example, a firm that emphasizes car production will devote less of its resources to producing tractors. In Fig. 7.11, curve O_2 lies twice as far from the origin as curve O_1 , signifying that this firm's production process exhibits constant returns to scale in the production of both commodities.

If curve O_1 were a straight line, joint production would entail no gains (or losses). One smaller company specializing in cars and another in tractors would generate the same output as a single company producing both. However, the

product transformation curve
Curve showing the various combinations of two different outputs (products) that can be produced with a given set of inputs.

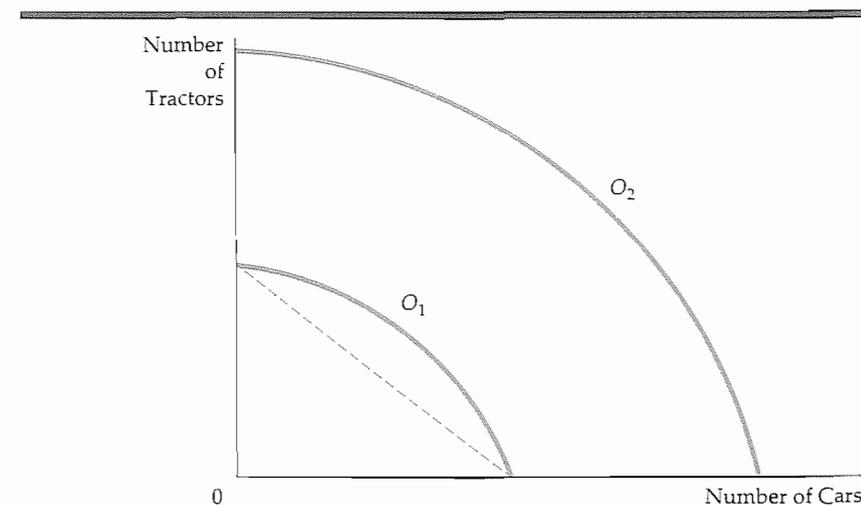


FIGURE 7.11 Product Transformation Curve

The product transformation curve describes the different combinations of two outputs that can be produced with a fixed amount of production inputs. The product transformation curves O_1 and O_2 are bowed out (or concave) because there are economies of scope in production.

product transformation curve is bowed outward (or *concave*) because joint production usually has advantages that enable a single company to produce more cars and tractors with the same resources than two companies producing each product separately. These production advantages involve the joint sharing of inputs. A single management, for example, is often able to schedule and organize production and to handle accounting and financial aspects more effectively than could separate managements.

Economies and Diseconomies of Scope

In general, **economies of scope** are present when the joint output of a single firm is greater than the output that could be achieved by two different firms each producing a single product (with equivalent production inputs allocated between the two firms). If a firm's joint output is *less* than that which could be achieved by separate firms, then its production process involves **diseconomies of scope**. This possibility could occur if the production of one product somehow conflicted with the production of the second.

There is no direct relationship between economies of scale and economies of scope. A two-output firm can enjoy economies of scope even if its production process involves diseconomies of scale. Suppose, for example, that manufacturing flutes and piccolos jointly is cheaper than producing both separately. Yet the production process involves highly skilled labor and is most effective if undertaken on a small scale. Likewise, a joint-product firm can have economies of scale for each individual product yet not enjoy economies of scope. Imagine, for example, a large conglomerate that owns several firms that produce efficiently on a large scale but that do not take advantage of economies of scope because they are administered separately.

economies of scope Joint output of a single firm is greater than output that could be achieved by two different firms when each produces a single product.

diseconomies of scope Joint output of a single firm is less than could be achieved by separate firms when each produces a single product.

The Degree of Economies of Scope

The extent to which there are economies of scope can also be determined by studying a firm's costs. If a combination of inputs used by one firm generates more output than two independent firms would produce, then it costs less for a single firm to produce both products than it would cost the independent firms. To measure the degree to which there are economies of scope, we should ask what percentage of the cost of production is saved when two (or more) products are produced jointly rather than individually. Equation (7.7) gives the **degree of economies of scope (SC)** that measures this savings in cost:

$$SC = \frac{C(Q_1) + C(Q_2) - C(Q_1, Q_2)}{C(Q_1, Q_2)} \quad (7.7)$$

$C(Q_1)$ represents the cost of producing output Q_1 , $C(Q_2)$ the cost of producing output Q_2 , and $C(Q_1, Q_2)$ the joint cost of producing both outputs. When the physical units of output can be added, as in the car-tractor example, the expression becomes $C(Q_1 + Q_2)$. With economies of scope, the joint cost is less than the sum of the individual costs. Thus, SC is greater than 0. With diseconomies of scope, SC is negative. In general, the larger the value of SC, the greater the economies of scope.

degree of economies of scope (SC) Percentage of cost savings resulting when two or more products are produced jointly rather than individually.

EXAMPLE 7.5 Economies of Scope in the Trucking Industry

Suppose that you are managing a trucking firm that hauls loads of different sizes between cities.⁹ In the trucking business, several related but distinct products can be offered, depending on the size of the load and the length of the haul. First, any load, small or large, can be taken directly from one location to another without intermediate stops. Second, a load can be combined with other loads (which may go between different locations) and eventually be shipped indirectly from its origin to the appropriate destination. Each type of load, partial or full, may involve different lengths of haul.

This raises questions about both economies of scale and economies of scope. The scale question is whether large-scale, direct hauls are cheaper and more profitable than individual hauls by small truckers. The scope question is whether a large trucking firm enjoys cost advantages in operating both direct quick hauls and indirect, slower (but less expensive) hauls. Central planning and organization of routes could provide for economies of scope. The key to the presence of economies of scale is the fact that the organization of routes and the types of hauls we have described can be accomplished more efficiently when many hauls are involved. In such cases, a firm is more likely to be able to schedule hauls in which most truckloads are full rather than half-full.

Studies of the trucking industry show that economies of scope are present. For example, one analysis of 105 trucking firms looked at four distinct outputs: (1) short hauls with partial loads, (2) intermediate hauls with partial loads, (3) long hauls with partial loads, and (4) hauls with total loads. The results indicate that the degree of economies of scope SC was 1.576 for a reasonably large firm. However, the degree of economies of scope falls to 0.104 when the firm becomes very large. Because large firms carry sufficiently large truckloads, there is usually no advantage to stopping at an intermediate terminal to fill a partial load. A direct trip from the origin to the destination is sufficient. Apparently, however, because other disadvantages are associated with the management of very large firms, the economies of scope get smaller as the firm gets bigger. In any event, the ability to combine partial loads at an intermediate location lowers the firm's costs and increases its profitability.

The study suggests, therefore, that to compete in the trucking industry a firm must be large enough to be able to combine loads at intermediate stopping points.

*7.6 Dynamic Changes in Costs— The Learning Curve

Our discussion thus far has suggested one reason a large firm may have a lower long-run average cost than a small firm: increasing returns to scale in production. It is tempting to conclude that firms which enjoy lower average cost over

⁹ This example is based on Judy S. Wang Chiang and Ann F. Friedlaender, "Truck Technology and Efficient Market Structure," *Review of Economics and Statistics* 67 (1985): 250–58.

time are growing firms with increasing returns to scale. But this need not be true. In some firms, long-run average cost may decline over time because workers and managers absorb new technological information as they become more experienced at their jobs.

As management and labor gain experience with production, the firm's marginal and average costs of producing a given level of output fall for four reasons:

1. Workers often take longer to accomplish a given task the first few times they do it. As they become more adept, their speed increases.
2. Managers learn to schedule the production process more effectively, from the flow of materials to the organization of the manufacturing itself.
3. Engineers who are initially cautious in their product designs may gain enough experience to be able to allow for tolerances in design that save cost without increasing defects. Better and more specialized tools and plant organization may also lower cost.
4. Suppliers of materials may learn how to process materials required more effectively and may pass on some of this advantage in the form of lower materials cost.

As a consequence, a firm "learns" over time as cumulative output increases. Managers can use this learning process to help plan production and forecast future costs. Figure 7.12 illustrates this process in the form of a **learning curve**—a curve that describes the relationship between a firm's cumulative output and the amount of inputs needed to produce each unit of output.

learning curve Graph relating amount of inputs needed by a firm to produce each unit of output to its cumulative output.

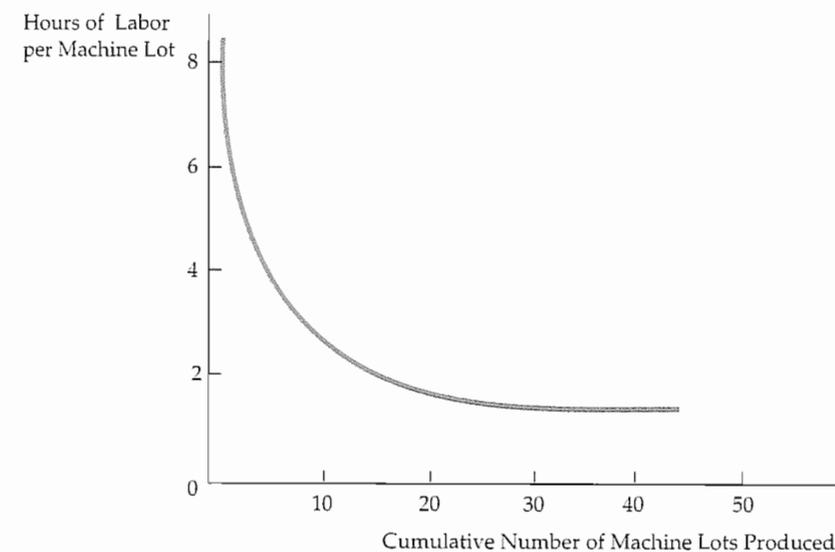


FIGURE 7.12 The Learning Curve

A firm's production cost may fall over time as managers and workers become more experienced and more effective at using the available plant and equipment. The learning curve shows the extent to which hours of labor needed per unit of output fall as the cumulative output increases.

Graphing the Learning Curve

Figure 7.12 shows a learning curve for the production of machine tools. The horizontal axis measures the *cumulative* number of lots of machine tools (groups of approximately 40) that the firm has produced. The vertical axis shows the number of hours of labor needed to produce each lot. Labor input per unit of output directly affects the production cost because the fewer the hours of labor needed, the lower the marginal and average cost of production.

The learning curve in Figure 7.12 is based on the relationship

$$L = A + BN^{-\beta} \tag{7.8}$$

where N is the cumulative units of output produced and L the labor input per unit of output. A , B , and β are constants, with A and B positive, and β between 0 and 1. When N is equal to 1, L is equal to $A + B$, so that $A + B$ measures the labor input required to produce the first unit of output. When β equals 0, labor input per unit of output remains the same as the cumulative level of output increases; there is no learning. When β is positive and N gets larger and larger, L becomes arbitrarily close to A . A , therefore, represents the minimum labor input per unit of output after all learning has taken place.

The larger is β , the more important is the learning effect. With β equal to 0.5, for example, the labor input per unit of output falls proportionally to the square root of the cumulative output. This degree of learning can substantially reduce the firm's production costs as the firm becomes more experienced.

In this machine tool example, the value of β is 0.31. For this particular learning curve, every doubling in cumulative output causes the input requirement (less the minimum attainable input requirement) to fall by about 20 percent.¹⁰ As Figure 7.12 shows, the learning curve drops sharply as the cumulative number of lots increases to about 20. Beyond an output of 20 lots, the cost savings are relatively small.

Learning versus Economies of Scale

Once the firm has produced 20 or more machine lots, the entire effect of the learning curve would be complete, and we could use the usual analysis of cost. If, however, the production process were relatively new, relatively high cost at low levels of output (and relatively low cost at higher levels) would indicate learning effects, not economies of scale. With learning, the cost of production for a mature firm is relatively low regardless of the scale of the firm's operation. If a firm that produces machine tools in lots knows that it enjoys economies of scale, it should produce its machines in very large lots to take advantage of the lower cost associated with size. If there is a learning curve, the firm can lower its cost by scheduling the production of many lots regardless of the individual lot size.

Figure 7.13 shows this phenomenon. AC_1 represents the long-run average cost of production of a firm that enjoys economies of scale in production. Thus the change in production from A to B along AC_1 leads to lower cost due to economies of scale. However, the move from A on AC_1 to C on AC_2 leads to lower cost due to learning, which shifts the average cost curve downward.

¹⁰Because $(L - A) = BN^{-\beta}$, we can check that $0.8(L - A)$ is approximately equal to $B(2N)^{-\beta}$.

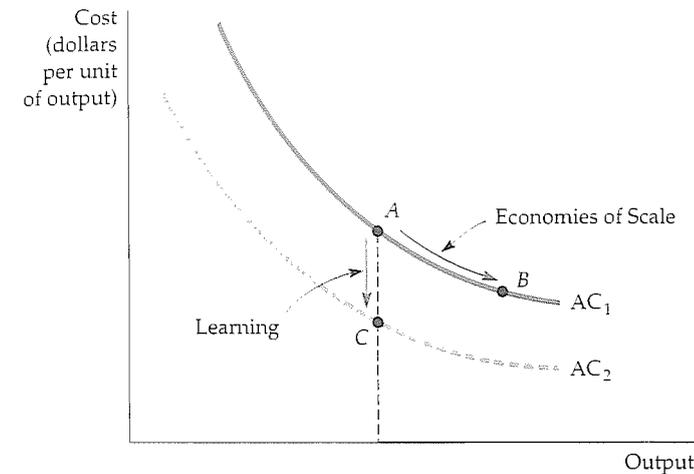


FIGURE 7.13 Economies of Scale versus Learning

A firm's average cost of production can decline over time because of growth of sales when increasing returns are present (a move from A to B on curve AC_1), or it can decline because there is a learning curve (a move from A on curve AC_1 to C on curve AC_2).

The learning curve is crucial for a firm that wants to predict the cost of producing a new product. Suppose, for example, that a firm producing machine tools knows that its labor requirement per machine for the first 10 machines is 1.0, the minimum labor requirement A is equal to zero, and β is approximately equal to 0.32. Table 7.3 calculates the total labor requirement for producing 80 machines.

Because there is a learning curve, the per-unit labor requirement falls with increased production. As a result, the total labor requirement for producing more and more output increases in smaller and smaller increments. Therefore, a

TABLE 7.3 Predicting the Labor Requirements of Producing a Given Output

CUMULATIVE OUTPUT (N)	PER-UNIT LABOR REQUIREMENT FOR EACH 10 UNITS OF OUTPUT (L)*	TOTAL LABOR REQUIREMENT
10	1.00	10.0
20	.80	18.0 (10.0 + 8.0)
30	.70	25.0 (18.0 + 7.0)
40	.64	31.4 (25.0 + 6.4)
50	.60	37.4 (31.4 + 6.0)
60	.56	43.0 (37.4 + 5.6)
70	.53	48.3 (43.0 + 5.3)
80	.51	53.4 (48.3 + 5.1)

*The numbers in this column were calculated from the equation $\log(L) = -0.322 \log(N/10)$, where L is the unit labor input and N is cumulative output.

firm looking only at the high initial labor requirement will obtain an overly pessimistic view of the business. Suppose the firm plans to be in business for a long time, producing 10 units per year. Suppose the total labor requirement for the first year's production is 10. In the first year of production, the firm's cost will be high as it learns the business. But once the learning effect has taken place, production costs will fall. After 8 years, the labor required to produce 10 units will be only 5.1, and per-unit cost will be roughly half what it was in the first year of production. Thus the learning curve can be important for a firm deciding whether it is profitable to enter an industry.

EXAMPLE 7.6 The Learning Curve in Practice

Suppose that as the manager of a firm that has just entered the chemical processing industry, you face the following problem: Should you produce a relatively low level of output and sell at a high price, or should you price your product lower and increase your rate of sales? The second alternative is appealing if there is a learning curve in this industry. In that case, the increased volume will lower your average production costs over time and increase the firm's profitability.

To decide what to do, you can examine the available statistical evidence that distinguishes the components of the learning curve (learning new processes by labor, engineering improvements, etc.) from increasing returns to scale. For example, a study of 37 chemical products reveals that cost reductions in the chemical processing industry are directly tied to the growth of cumulative industry output, to investment in improved capital equipment, and, to a lesser extent, to economies of scale.¹¹ In fact, for the entire sample of chemical products, average costs of production fall at 5.5 percent per year. The study reveals that for each doubling of plant scale, the average cost of production falls by 11 percent. For each doubling of cumulative output, however, the average cost of production falls by 27 percent. The evidence shows clearly that learning effects are more important than economies of scale in the chemical processing industry.¹²

The learning curve has also been shown to be important in the semiconductor industry. A study of seven generations of dynamic random-access memory (DRAM) semiconductors from 1974 to 1992 found that the learning rates averaged about 20 percent; thus a 10-percent increase in cumulative production

¹¹ The study was conducted by Marvin Lieberman, "The Learning Curve and Pricing in the Chemical Processing Industries," *RAND Journal of Economics* 15 (1984): 213–28.

¹² The author used the average cost AC of the chemical products, the cumulative industry output X, and the average scale of a production plant Z. He then estimated the relationship $\log(AC) = -0.387 \log(X) - 0.173 \log(Z)$. The -0.387 coefficient on cumulative output tells us that for every 1-percent increase in cumulative output, average cost decreases 0.387 percent. The -0.173 coefficient on plant size tells us that for every 1-percent increase in plant size, cost decreases 0.173 percent.

By interpreting the two coefficients in light of the output and plant-size variables, we can allocate about 15 percent of the cost reduction to increases in the average scale of plants and 85 percent to increases in cumulative industry output. Suppose plant scale doubled while cumulative output increased by a factor of 5 during the study. In that case, costs would fall by 11 percent from the increased scale and by 62 percent from the increase in cumulative output.

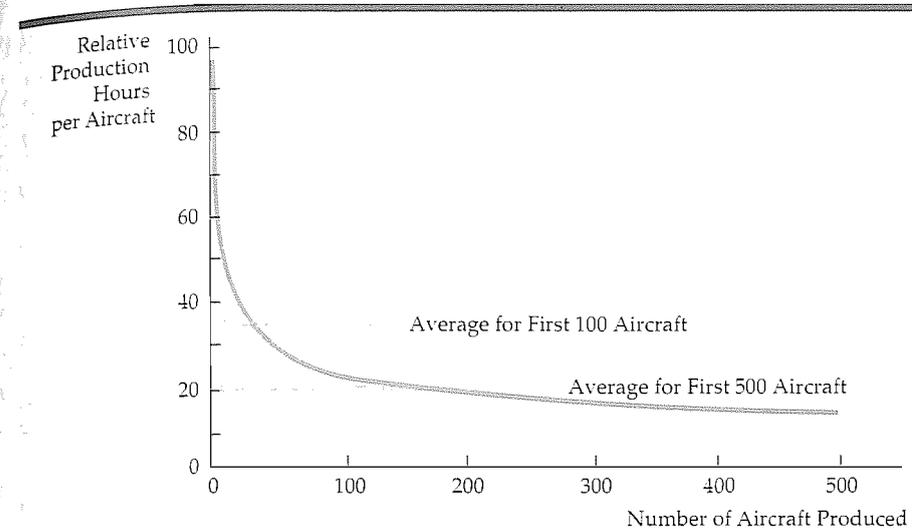


FIGURE 7.14 Learning Curve for Airbus Industrie

The learning curve relates the labor requirement per aircraft to the cumulative number of aircraft produced. As the production process becomes better organized and workers gain familiarity with their jobs, labor requirements fall dramatically.

would lead to a 2-percent decrease in cost.¹³ The study also compared learning by firms in Japan to firms in the United States and found that there was no distinguishable difference in the speed of learning.

Another example is the aircraft industry, where studies have found learning rates that are as high as 40 percent. This is illustrated in Figure 7.14, which shows the labor requirements for production of aircraft by Airbus Industrie. Observe that the first 10 or 20 airplanes require far more labor to produce than the hundredth or two hundredth airplane. Also note how the learning curve flattens out after a certain point; in this case, nearly all learning is complete after 200 airplanes have been built.

Learning curve effects can be important in determining the shape of long-run cost curves and can thus help guide management decisions. Managers can use learning curve information to decide whether a production operation is profitable and, if it is, how to plan how large the plant operation and the volume of cumulative output need be to generate a positive cash flow.

*7.7 Estimating and Predicting Cost

A business that is expanding or contracting its operation must predict how costs will change as output changes. Estimates of future costs can be obtained from a **cost function**, which relates the cost of production to the level of output and other variables that the firm can control.

cost function Function relating cost of production to level of output and other variables that the firm can control.

¹³ The study was conducted by D. A. Irwin and P. J. Klenow, "Learning-by-Doing Spillovers in the Semiconductor Industry," *Journal of Political Economy* 102 (December 1994): 1200–27.

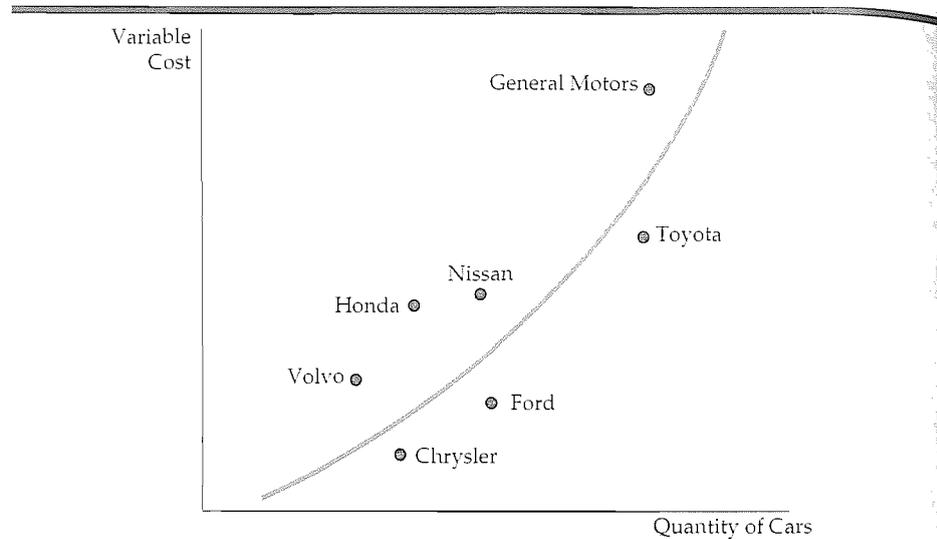


FIGURE 7.15 Total Cost Curve for the Automobile Industry

An empirical estimate of the total cost curve can be obtained by using data for individual firms in an industry. The total cost curve for automobile production is obtained by determining statistically the curve that best fits the points that relate the output of each firm to the total cost of production.

Suppose we wanted to characterize the short-run cost of production in the automobile industry. We could obtain data on the number of automobiles Q produced by each car company and relate this information to the variable cost of production VC . The use of variable cost, rather than total cost, avoids the problem of trying to allocate the fixed cost of a multiproduct firm's production process to the particular product being studied.¹⁴

Figure 7.15 shows a typical pattern of cost and output data. Each point on the graph relates the output of an auto company to that company's variable cost of production. To predict cost accurately, we must determine the underlying relationship between variable cost and output. Then, if a company expands its production, we can calculate what the associated cost is likely to be. The curve in the figure is drawn with this in mind—it provides a reasonably close fit to the cost data. (Typically, least-squares regression analysis would be used to fit the curve to the data.) But what shape is the most appropriate, and how do we represent that shape algebraically?

Here is one cost function that we might choose:

$$VC = \beta Q \tag{7.9}$$

Although easy to use, this *linear* relationship between cost and output is applicable only if marginal cost is constant.¹⁵ For every unit increase in output, variable cost increases by β ; marginal cost is thus constant and equal to β .

¹⁴ If an additional piece of equipment is needed as output increases, then the annual rental cost of the equipment should be counted as a variable cost. If, however, the same machine can be used at all output levels, its cost is fixed and should not be included.

¹⁵ In statistical cost analyses, other variables might be added to the cost function to account for differences in input costs, production processes, production mix, etc., among firms.

Least-squares regression is explained in the appendix to this book.

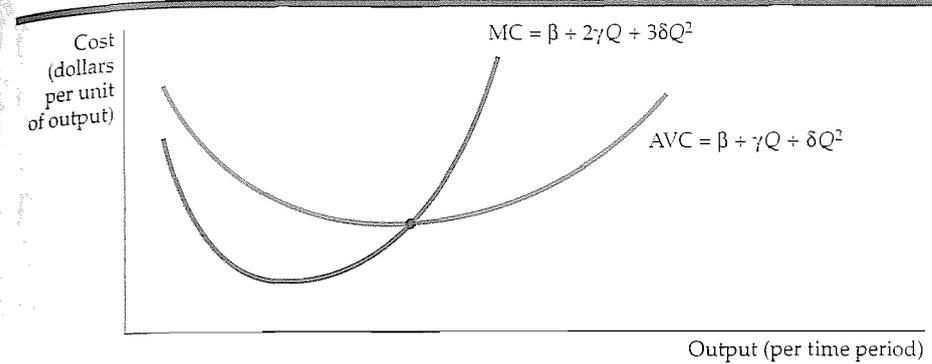


FIGURE 7.16 Cubic Cost Function

A cubic cost function implies that the average and the marginal cost curves are U-shaped.

If we wish to allow for a U-shaped average cost curve and a marginal cost that is not constant, we must use a more complex cost function. One possibility is the *quadratic* cost function, which relates variable cost to output and output squared:

$$VC = \beta Q + \gamma Q^2 \tag{7.10}$$

This function implies a straight-line marginal cost curve of the form $MC = \beta + 2\gamma Q$.¹⁶ Marginal cost increases with output if γ is positive and decreases with output if γ is negative.

If the marginal cost curve is not linear, we might use a *cubic* cost function:

$$VC = \beta Q + \gamma Q^2 + \delta Q^3 \tag{7.11}$$

Figure 7.16 shows this cubic cost function. It implies U-shaped marginal as well as average cost curves.

Cost functions can be difficult to measure for several reasons. First, output data often represent an aggregate of different types of products. The automobiles produced by General Motors, for example, involve different models of cars. Second, cost data are often obtained directly from accounting information that fails to reflect opportunity costs. Third, allocating maintenance and other plant costs to a particular product is difficult when the firm is a conglomerate that produces more than one product line.

Cost Functions and the Measurement of Scale Economies

Recall that the cost-output elasticity E_C is less than one when there are economies of scale and greater than one when there are diseconomies of scale. The *scale economies index (SCI)* provides an index of whether or not there are scale economies. SCI is defined as follows:

$$SCI = 1 - E_C \tag{7.12}$$

¹⁶ Short-run marginal cost is given by $\Delta VC/\Delta Q = \beta + \gamma\Delta(Q^2)$. But $\Delta(Q^2)/\Delta Q = 2Q$. (Check this by using calculus or by numerical example.) Therefore, $MC = \beta + 2\gamma Q$.

When $E_C = 1$, $SCI = 0$ and there are no economies or diseconomies of scale. When E_C is greater than one, SCI is negative and there are diseconomies of scale. Finally, when E_C is less than 1, SCI is positive and there are economies of scale.

EXAMPLE 7.7 Cost Functions for Electric Power

In 1955, consumers bought 369 billion kilowatt-hours (kwh) of electricity; in 1970 they bought 1083 billion. Because there were fewer electric utilities in 1970, the output per firm had increased substantially. Was this increase due to economies of scale or other factors? If it was the result of economies of scale, it would be economically inefficient for regulators to “break up” electric utility monopolies.

An interesting study of scale economies was based on the years 1955 and 1970 for investor-owned utilities with more than \$1 million in revenues.¹⁷ The cost of electric power was estimated by using a cost function that is somewhat more sophisticated than the quadratic and cubic functions discussed earlier.¹⁸ Table 7.4 shows the resulting estimates of the scale economies index. The results are based on a classification of all utilities into five size categories, with the median output (measured in kilowatt-hours) in each category listed.

The positive values of SCI tells us that all sizes of firms had some economies of scale in 1955. However, the magnitude of the economies of scale diminishes as firm size increases. The average cost curve associated with the 1955 study is drawn in Figure 7.17 and labeled 1955. The point of minimum average cost occurs at point A at an output of approximately 20 billion kilowatts. Because there were no firms of this size in 1955, no firm had exhausted the opportunity for returns to scale in production. Note, however, that the average cost curve is relatively flat from an output of 9 billion kilowatts and higher, a range in which 7 of 124 firms produced.

When the same cost functions were estimated with 1970 data, the cost curve labeled 1970 in Figure 7.17 was the result. The graph shows clearly that the average costs of production fell from 1955 to 1970. (The data are in real 1970 dollars.) But the flat part of the curve now begins at about 15 billion kwh. By 1970, 24 of 80 firms were producing in this range. Thus many more firms were operating in the flat portion of the average cost curve in which economies of scale are not an important phenomenon. More important, most of the firms were producing in a portion of the 1970 cost curve that was flatter than their point of operation on the 1955 curve. (Five firms were at points of diseconomies

TABLE 7.4 Scale Economies in the Electric Power Industry

Output (million kwh)	43	338	1109	2226	5819
Value of SCI , 1955	.41	.26	.16	.10	.04

¹⁷ This example is based on Laurits Christensen and William H. Greene, “Economies of Scale in U.S. Electric Power Generation,” *Journal of Political Economy* 84 (1976): 655–76.

¹⁸ The translog cost function used in this study provides a more general functional relationship than any of those we have discussed.

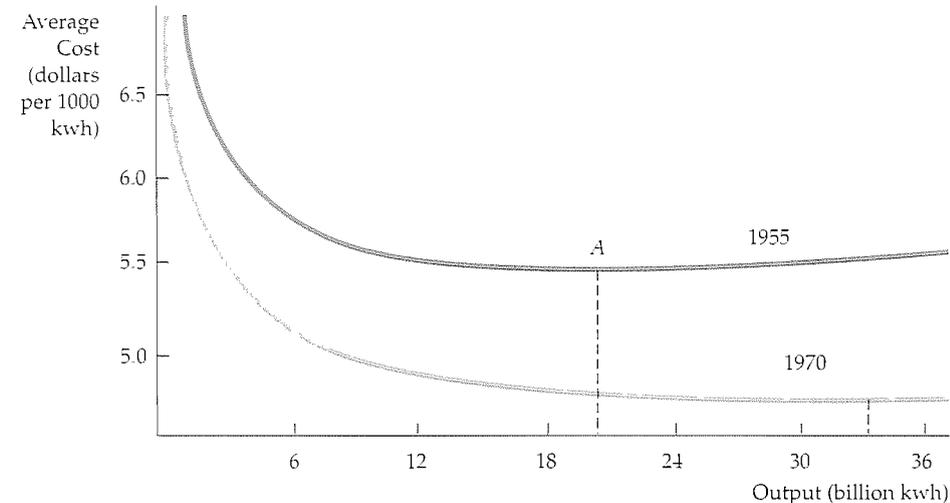


FIGURE 7.17 Average Cost of Production in the Electric Power Industry

The average cost of electric power in 1955 achieved a minimum at approximately 20 billion kilowatt-hours. By 1970 the average cost of production had fallen sharply and achieved a minimum at an output of more than 33 billion kilowatt-hours.

of scale: Consolidated Edison [$SCI = -0.003$], Detroit Edison [$SCI = -0.004$], Duke Power [$SCI = -0.012$], Commonwealth Edison [$SCI = -0.014$], and Southern [$SCI = -0.028$].) Thus unexploited scale economies were much smaller in 1970 than in 1955.

This cost function analysis makes it clear that the decline in the cost of producing electric power cannot be explained by the ability of larger firms to take advantage of economies of scale. Rather, improvements in technology unrelated to the scale of the firms’ operation and the decline in the real cost of energy inputs, such as coal and oil, are important reasons for the lower costs. The tendency toward lower average cost reflecting a movement to the right along an average cost curve is minimal compared with the effect of technological improvement.

EXAMPLE 7.8 A Cost Function for the Savings and Loan Industry

Understanding returns to scale in the savings and loan industry is important for regulators who must decide how savings and loans should be restructured in light of the failure of numerous institutions. In this regard, the empirical estimation of a long-run cost function can be useful.¹⁹

Data were collected for 86 savings and loan associations for 1975 and 1976 in a region that includes Idaho, Montana, Oregon, Utah, Washington, and Wyoming. Output is difficult to measure in this case because a savings and

¹⁹ This example builds on J. Holton Wilson, “A Note on Scale Economies in the Savings and Loan Industry,” *Business Economics* (January 1981): 45–49.

loan association provides a service to its customers, rather than a physical product. The output Q measure reported here (and used in other studies) is the total assets of each savings and loan association. In general, the larger the asset base of an association, the higher its profitability. Long-run average cost LAC is measured by average operating expense. Output and total operating costs are measured in hundreds of millions of dollars. Average operating costs are measured as a percentage of total assets.

A quadratic long-run average cost function was estimated for the year 1975, yielding the following relationship:

$$\text{LAC} = 2.38 - 0.6153Q + 0.0536Q^2$$

The estimated long-run average cost function is U-shaped and reaches its point of minimum average cost when the total assets of the savings and loan reach \$574 million.²⁰ (At this point the average operating expenses of the savings and loan are 0.61 percent of its total assets.) Because almost all savings and loans in the region being studied had substantially less than \$574 million in assets, the cost function analysis suggests that an expansion of savings and loans through either growth or mergers would be valuable.

How appropriate such a policy is cannot be fully evaluated here, however. To do so, we would need to take into account the possible social costs associated with the lessening of competition from growth or mergers, and we would need to assure ourselves that this particular cost function analysis accurately estimated the point of minimum average cost.

SUMMARY

1. Managers, investors, and economists must take into account the *opportunity cost* associated with the use of a firm's resources: the cost associated with the opportunities forgone when the firm uses its resources in its next best alternative.
2. A *sunk cost* is an expenditure that has been made and cannot be recovered. After it has been incurred, it should be ignored when making future economic decisions.
3. In the short run, one or more of the firm's inputs are fixed. Total cost can be divided into fixed cost and variable cost. A firm's *marginal cost* is the additional variable cost associated with each additional unit of output. The *average variable cost* is the total variable cost divided by the number of units of output.
4. In the short run, when not all inputs are variable, the presence of diminishing returns determines the shape of the cost curves. In particular, there is an inverse relationship between the marginal product of a

single variable input and the marginal cost of production. The average variable cost and average total cost curves are U-shaped. The short-run marginal cost curve increases beyond a certain point, and cuts both average cost curves from below at their minimum points.

5. In the long run, all inputs to the production process are variable. As a result, the choice of inputs depends both on the relative costs of the factors of production and on the extent to which the firm can substitute among inputs in its production process. The cost-minimizing input choice is made by finding the point of tangency between the isoquant representing the level of desired output and an isocost line.
6. The firm's expansion path shows how its cost-minimizing input choices vary as the scale or output of its operation increases. As a result, the expansion path provides useful information relevant for long-run planning decisions.

²⁰You can confirm this principle either by graphing the curve or by differentiating the average cost function with respect to Q , setting it equal to 0, and solving for Q .

7. The long-run average cost curve is the envelope of the firm's short-run average cost curves, and it reflects the presence or absence of returns to scale. When there are constant returns to scale and many plant sizes are possible, the long-run cost curve is horizontal; the envelope consists of the points of minimum short-run average cost. However, when there are increasing returns to scale initially and then decreasing returns to scale, the long-run average cost curve is U-shaped, and the envelope does not include all points of minimum short-run average cost.
8. A firm enjoys *economies of scale* when it can double its output at less than twice the cost. Correspondingly, there are *diseconomies of scale* when a doubling of output requires more than twice the cost. Scale economies and diseconomies apply even when input proportions are variable; returns to scale applies only when input proportions are fixed.
9. When a firm produces two (or more) outputs, it is important to note whether there are *economies of scope*

in production. Economies of scope arise when the firm can produce any combination of the two outputs more cheaply than could two independent firms that each produced a single product. The degree of economies of scope is measured by the percentage reduction in cost when one firm produces two products relative to the cost of producing them individually.

10. A firm's average cost of production can fall over time if the firm "learns" how to produce more effectively. The *learning curve* shows how much the input needed to produce a given output falls as the cumulative output of the firm increases.
11. Cost functions relate the cost of production to the firm's level of output. The functions can be measured in both the short run and the long run by using either data for firms in an industry at a given time or data for an industry over time. A number of functional relationships, including linear, quadratic, and cubic, can be used to represent cost functions.

QUESTIONS FOR REVIEW

1. A firm pays its accountant an annual retainer of \$10,000. Is this an explicit or an implicit cost?
2. The owner of a small retail store does her own accounting work. How would you measure the opportunity cost of her work?
3. Suppose a chair manufacturer finds that the marginal rate of technical substitution of capital for labor in his production process is substantially greater than the ratio of the rental rate on machinery to the wage rate for assembly-line labor. How should he alter his use of capital and labor to minimize the cost of production?
4. Why are isocost lines straight lines?
5. If the marginal cost of production is increasing, do you know whether the average variable cost is increasing or decreasing? Explain.
6. If the marginal cost of production is greater than the average variable cost, do you know whether the average variable cost is increasing or decreasing? Explain.
7. If a firm's average cost curves are U-shaped, why does its average variable cost curve achieve its minimum at a lower level of output than the average total cost curve?
8. If a firm enjoys increasing returns to scale up to a certain output level, and then constant returns to scale, what can you say about the shape of its long-run average cost curve?
9. How does a change in the price of one input change a firm's long-run expansion path?
10. Distinguish between economies of scale and economies of scope. Why can one be present without the other?

EXERCISES

1. Assume a computer firm's marginal costs of production are constant at \$1000 per computer. However, the fixed costs of production are equal to \$10,000.
 - a. Calculate the firm's average variable cost and average total cost curves.
 - b. If the firm wanted to minimize the average total cost of production, would it choose to be very large or very small? Explain.
2. If a firm hires a currently unemployed worker, the opportunity cost of utilizing the worker's service is zero. Is this true? Discuss.
3. a. Suppose a firm must pay an annual franchise fee or tax, which is a fixed sum, independent of whether it produces any output. How does this tax affect the firm's fixed, marginal, and average costs?

b. Now suppose the firm is charged a tax that is proportional to the number of items it produces. Again, how does this tax affect the firm's fixed, marginal, and average costs?

4. Several years ago *Business Week* reported the following:

During an auto sales slump, GM, Ford, and Chrysler decided it was cheaper to sell cars to rental companies at a loss than to lay off workers. That's because closing and reopening plants is expensive, partly because the automakers' current union contracts obligate them to pay many workers even if they're not working.

When the article discusses selling cars "at a loss," is it referring to accounting profit or economic profit? How will the two differ in this case? Explain briefly.

5. A chair manufacturer hires its assembly-line labor for \$22 an hour and calculates that the rental cost of its machinery is \$110 per hour. Suppose that a chair can be produced using 4 hours of labor or machinery in any combination. If the firm is currently using 3 hours of labor for each hour of machine time, is it minimizing its costs of production? If so, why? If not, how can it improve the situation?
6. Suppose the economy takes a downturn; labor costs fall by 50 percent and are expected to stay at that level for a long time. Show graphically how this change in the relative price of labor and capital affects a firm's expansion path.
7. You are in charge of cost control in a large metropolitan transit district. A consultant comes to you with the following report:

Our research has shown that the cost of running a bus for each trip down its line is \$30 regardless of the number of passengers it carries. Each bus can carry 50 people. At rush hour, when the buses are full, the average cost per passenger is 60 cents. However, during off-peak hours, average ridership falls to 18 people and average cost soars to \$1.67 per passenger. As a result, we should encourage more rush-hour business when costs are cheaper and discourage off-peak business when costs are higher.

Do you follow the consultant's advice? Discuss.

8. An oil refinery consists of different pieces of processing equipment, each of which differs in its ability to break down heavy sulfurized crude oil into final products. The refinery process is such that the marginal cost of producing gasoline is constant up to a point as crude oil is put through a basic distilling unit. However, as the unit fills up, the firm finds that in the short run, the amount of crude oil that can be processed is limited. The marginal cost of producing gasoline is also

constant up to a capacity limit when crude oil is put through a more sophisticated hydrocracking unit. Graph the marginal cost of gasoline production when a basic distilling unit and a hydrocracker are used.

9. You manage a plant that mass produces engines by teams of workers using assembly machines. The technology is summarized by the production function

$$Q = 4KL$$

where Q is the number of engines per week, K the number of assembly machines, and L the number of labor teams. Each assembly machine rents for $r = \$12,000$ per week and each team costs $w = \$3,000$ per week. Engine costs are given by the cost of labor teams and machines, plus \$2,000 per engine for raw materials. Your plant has a fixed installation of 10 assembly machines as part of its design.

- a. What is the cost function for your plant—namely, how much will it cost to produce Q engines? What are average and marginal costs for producing Q engines? How do average costs vary with output?
- b. How many teams are required for producing 80 engines? What is the average cost per engine?
- c. You are asked to make recommendations for the design of a new production facility. What would you suggest? In particular, what capital/labor (K/L) ratio should the new plant accommodate? If lower average cost were your only criterion, should you suggest that the new plant have more or less production capacity than the plant you currently manage?
- *10. A computer company's cost function relates its average cost of production AC to its cumulative output in thousands of computers CQ . Its plant size in terms of thousands of computers produced per year Q , within the production range of 10,000 to 50,000 computers, is given by

$$AC = 10 - 0.1CQ + 0.3Q$$

- a. Is there a learning curve effect?
- b. Are there increasing or decreasing returns to scale?
- c. During its existence, the firm has produced a total of 40,000 computers and is producing 10,000 computers this year. Next year it plans to increase its production to 12,000 computers. Will its average cost of production increase or decrease? Explain.
11. The total short-run cost function of a company is given by the equation $C = 190 + 53Q$, where C is the total cost and Q is the total quantity of output, both measured in tens of thousands.
- a. What is the company's fixed cost?
- b. If the company produces 100,000 units, what is its average variable cost?
- c. What is its marginal cost *per unit* produced?
- d. What is its average fixed cost?

- e. Suppose the company borrows money and expands its factory. Its fixed cost rises by \$50,000, but its variable cost falls to \$45,000 per 10,000 units. The interest cost (I) also enters into the equation. Each one-point increase in the interest rate raises costs by \$30,000. Write the new cost equation.

- *12. Suppose the long-run total cost function for an industry is given by the cubic equation $TC = a + bQ + cQ^2 + dQ^3$. Show (using calculus) that this total cost function is consistent with a U-shaped average cost curve for at least some values of the parameters a, b, c, d .

- *13. A computer company produces hardware and software using the same plant and labor. The total cost of producing computer processing units H and software programs S is given by

$$TC = aH + bS - cHS$$

where a, b , and c are positive. Is this total cost function consistent with the presence of economies or diseconomies of scale? With economies or diseconomies of scope?

APPENDIX TO CHAPTER 7

Production and Cost Theory— A Mathematical Treatment

This appendix presents a mathematical treatment of the basics of production and cost theory. As in the appendix to Chapter 4, we use the method of Lagrange multipliers to solve the firm's cost-minimizing problem.

Cost Minimization

The theory of the firm relies on the assumption that firms choose inputs to the production process that minimize the cost of producing output. If there are two inputs, capital K and labor L , the production function $F(K, L)$ describes the maximum output that can be produced for every possible combination of inputs. We assume that each of the factors in the production process has positive but decreasing marginal products. Writing the marginal product of capital as $MP_K(K, L) = \partial F(K, L)/\partial K$, we assume that $MP_K(K, L) > 0$ and $\partial MP_K(K, L)/\partial K < 0$. Similarly, if the marginal product of labor is given by $MP_L(K, L) = \partial F(K, L)/\partial L$, we assume that $MP_L(K, L) > 0$ and $\partial MP_L(K, L)/\partial L < 0$.

A competitive firm takes the prices of both labor w and capital r as given. Then the cost-minimization problem can be written as

$$\text{Minimize } C = wL + rK \quad (\text{A7.1})$$

subject to the constraint that a fixed output Q_0 be produced:

$$F(K, L) = Q_0 \quad (\text{A7.2})$$

C represents the cost of producing the fixed level of output Q_0 .

To determine the firm's demand for capital and labor inputs, we choose the values of K and L that minimize (A7.1) subject to (A7.2). We can solve this constrained optimization problem in three steps using the method discussed in the Appendix to Chapter 4:

- **Step 1.** Set up the Lagrangian, which is the sum of two components: the cost of production (to be minimized) and the Lagrange multiplier λ times the output constraint faced by the firm:

$$\Phi = wL + rK - \lambda[F(K, L) - Q_0] \quad (\text{A7.3})$$

- **Step 2.** Differentiate the Lagrangian with respect to K , L , and λ . Then equate the resulting derivatives to zero to obtain the necessary conditions for a minimum:¹

$$\begin{aligned} \partial\Phi/\partial K &= r - \lambda MP_K(K, L) = 0 \\ \partial\Phi/\partial L &= w - \lambda MP_L(K, L) = 0 \\ \partial\Phi/\partial \lambda &= F(K, L) - Q_0 = 0 \end{aligned} \quad (\text{A7.4})$$

¹ These conditions are necessary for a solution involving positive amounts of both inputs.

- **Step 3.** In general, these equations can be solved to obtain the optimizing values of L , K , and λ . It is particularly instructive to combine the first two conditions in (A7.4), to obtain

$$MP_K(K, L)/r = MP_L(K, L)/w \quad (\text{A7.5})$$

Equation (A7.5) tells us that if the firm is minimizing costs it will choose its factor inputs to equate the ratio of the marginal product of each factor divided by its price. To see that this makes sense, suppose MP_K/r were greater than MP_L/w . Then the firm could reduce its cost while still producing the same output by using more capital and less labor.

Finally, we can combine the first two conditions of (A7.4) in a different way to evaluate the Lagrange multiplier:

$$\lambda = r/MP_K(K, L) = w/MP_L(K, L) \quad (\text{A7.6})$$

Suppose output increases by one unit. Because the marginal product of capital measures the extra output associated with an additional input of capital, $1/MP_K(K, L)$ measures the extra capital needed to produce one unit of output. Therefore, $r/MP_K(K, L)$ measures the additional input cost of producing an additional unit of output by increasing capital. Likewise, $w/MP_L(K, L)$ measures the additional cost of producing a unit of output using additional labor as an input. In both cases, the Lagrange multiplier is equal to the marginal cost of production, because it tells us how much the cost increases if the amount is increased by one unit.

Marginal Rate of Technical Substitution

Recall that an *isoquant* is a curve that represents the set of all input combinations that give the firm the same level of output—say, Q^* . Thus the condition that $F(K, L) = Q^*$ represents a production isoquant. As input combinations are changed along an isoquant, the change in output, given by the total derivative of $F(K, L)$ equals zero (i.e., $dQ = 0$). Thus

$$MP_K(K, L)dK + MP_L(K, L)dL = dQ = 0 \quad (\text{A7.7})$$

It follows by rearrangement that

$$-dK/dL = MRTS_{LK} = MP_L(K, L)/MP_K(K, L) \quad (\text{A7.8})$$

where $MRTS_{LK}$ is the firm's marginal rate of technical substitution between labor and capital.

Now, rewrite the condition given by (A7.5) to get

$$MP_L(K, L)/MP_K(K, L) = w/r \quad (\text{A7.9})$$

Because the left side of (A7.8) represents the negative of the slope of the isoquant, it follows that at the point of tangency of the isoquant and the isocost line, the firm's marginal rate of technical substitution (which trades off inputs while keeping output constant) is equal to the ratio of the input prices (which represents the slope of the firm's isocost line).

We can look at this result another way by rewriting (A7.9) again:

$$MP_L/w = MP_K/r \quad (\text{A7.10})$$

Equation (A7.10) tells us that the marginal products of all production inputs must be equal when these marginal products are adjusted by the unit cost of each input. If the cost-adjusted marginal products were not equal, the firm could change its inputs to produce the same output at a lower cost.

Duality in Production and Cost Theory

As in consumer theory, the firm's input decision has a dual nature. The optimum choice of K and L can be analyzed not only as the problem of choosing the lowest isocost line tangent to the production isoquant, but also as the problem of choosing the highest production isoquant tangent to a given isocost line. To verify this, consider the following dual producer problem:

$$\text{Maximize } F(K, L)$$

subject to the cost constraint that

$$wL + rK = C_0 \quad (\text{A7.11})$$

The corresponding Lagrangian is given by

$$\Phi = F(K, L) - \mu(wL + rK - C_0) \quad (\text{A7.12})$$

where μ is the Lagrange multiplier. The necessary conditions for output maximization are

$$\begin{aligned} MP_K(K, L) - \mu r &= 0 \\ MP_L(K, L) - \mu w &= 0 \\ wL + rK - C_0 &= 0 \end{aligned} \quad (\text{A7.13})$$

By solving the first two equations, we see that

$$MP_K(K, L)/r = MP_L(K, L)/w \quad (\text{A7.14})$$

which is identical to the condition that was necessary for cost minimization.

The Cobb-Douglas Cost and Production Functions

Given a specific production function $F(K, L)$, conditions (A7.13) and (A7.14) can be used to derive the *cost function* $C(Q)$. To understand this principle, let's work through the example of a **Cobb-Douglas production function**. This production function is

$$F(K, L) = AK^\alpha L^\beta$$

or, by taking the logs of both sides of the production function equation:

$$\log [F(K, L)] = \log A + \alpha \log K + \beta \log L$$

We assume that $\alpha < 1$ and $\beta < 1$, so that the firm has decreasing marginal products of labor and capital.² If $\alpha + \beta = 1$, the firm has *constant returns to scale*, because doubling K and L doubles F . If $\alpha + \beta > 1$, the firm has *increasing returns to scale*, and if $\alpha + \beta < 1$, it has *decreasing returns to scale*.

As an application, consider the carpet industry described in Example 6.4. The production of both small and large firms can be described by Cobb-Douglas production functions. For small firms, $\alpha = .77$ and $\beta = .23$. Because $\alpha + \beta = 1$, there is constant returns to scale. For larger firms, however, $\alpha = .83$ and $\beta = .22$. Thus $\alpha + \beta = 1.05$, and there is increasing returns to scale.

To find the amounts of capital and labor that the firm should utilize to minimize the cost of producing an output Q_0 , we first write the Lagrangian

$$\Phi = wL + rK - \lambda(AK^\alpha L^\beta - Q_0) \quad (\text{A7.15})$$

Differentiating with respect to L , K , and λ , and setting those derivatives equal to 0, we obtain

$$\partial\Phi/\partial L = w - \lambda(\beta AK^\alpha L^{\beta-1}) = 0 \quad (\text{A7.16})$$

$$\partial\Phi/\partial K = r - \lambda(\alpha AK^{\alpha-1} L^\beta) = 0 \quad (\text{A7.17})$$

$$\partial\Phi/\partial\lambda = AK^\alpha L^\beta - Q_0 = 0 \quad (\text{A7.18})$$

From equation (A7.16) we have

$$\lambda = w/\beta AK^\alpha L^{\beta-1} \quad (\text{A7.19})$$

Substituting this formula into equation (A7.17) gives us

$$r\beta AK^\alpha L^{\beta-1} = w\alpha AK^{\alpha-1} L^\beta \quad (\text{A7.20})$$

or

$$L = \beta r K / \alpha w \quad (\text{A7.21})$$

Now, use equation (A7.21) to eliminate L from equation (A7.18):

$$AK^\alpha \beta^\beta r^\beta K^\beta / \alpha^\beta w^\beta = Q_0 \quad (\text{A7.22})$$

Rewrite the new equation as

$$K^{\alpha+\beta} = (\alpha w / \beta r)^\beta Q_0 / A \quad (\text{A7.23})$$

or

$$K = [(\alpha w / \beta r)^{\beta/(\alpha+\beta)} (Q_0/A)^{1/(\alpha+\beta)}] \quad (\text{A7.24})$$

We have now determined the cost-minimizing quantity of capital. To determine the cost-minimizing quantity of labor, we simply substitute equation (A7.24) into equation (A7.21):

$$L = [(\beta r / \alpha w)^{\alpha/(\alpha+\beta)} (Q_0/A)^{1/(\alpha+\beta)}] \quad (\text{A7.25})$$

Note that if the wage rate w rises relative to the price of capital r , the firm will use more capital and less labor. Suppose that because of technological change, A

Cobb-Douglas production function Production function of the form $Q = AK^\alpha L^\beta$, where Q is the rate of output, K is the quantity of capital, and L is the quantity of labor, and where α and β are constants.

² For example, if the marginal product of labor is given by $MP_L = \partial[F(K, L)]/\partial L = \beta AK^\alpha L^{\beta-1}$, MP_L falls as L increases.

increases (so the firm can produce more output with the same inputs); in that case, both K and L will fall.

We have shown how cost-minimization subject to an output constraint can be used to determine the firm's optimal mix of capital and labor. Now we will determine the firm's cost function. The total cost of producing *any* output Q can be obtained by substituting equations (A7.24) for K and (A7.25) for L into the equation $C = wL + rK$. After some algebraic manipulation we find that

$$C = w^{\beta/(\alpha+\beta)} r^{\alpha/(\alpha+\beta)} \left[\left(\frac{\alpha}{\beta} \right)^{\beta/(\alpha+\beta)} + \left(\frac{\alpha}{\beta} \right)^{-\alpha/(\alpha+\beta)} \right] \left(\frac{Q}{A} \right)^{1/(\alpha+\beta)} \quad (\text{A7.26})$$

This *cost function* tells us (1) how the total cost of production increases as the level of output Q increases, and (2) how cost changes as input prices change. When $\alpha + \beta$ equals 1, equation (A7.26) simplifies to

$$C = w^{\beta} r^{\alpha} [(\alpha/\beta)^{\beta} + (\alpha/\beta)^{-\alpha}] (1/A) Q$$

In this case, therefore, cost will increase proportionately with output. As a result, the production process exhibits constant returns to scale. Likewise if $\alpha + \beta$ is greater than 1, there is decreasing returns to scale; if $\alpha + \beta$ is less than 1, there is increasing returns to scale.

Now consider the dual problem of maximizing the output that can be produced with the expenditure of C_0 dollars. We leave it to you to work through this problem for the Cobb-Douglas production function. You should be able to show that equations (A7.24) and (A7.25) describe the cost-minimizing input choices. To get you started, note that the Lagrangian for this dual problem is $\Phi = AK^{\alpha}L^{\beta} - \mu(wL + rK - C_0)$.

EXERCISES

- Of the following production functions, which exhibit increasing, constant, or decreasing returns to scale?
 - $F(K, L) = K^2L$
 - $F(K, L) = 10K + 5L$
 - $F(K, L) = (KL)^5$
- The production function for a product is given by $Q = 100KL$. If the price of capital is \$120 per day and the price of labor \$30 per day, what is the minimum cost of producing 1000 units of output?
- Suppose a production function is given by $F(K, L) = KL^2$; the price of capital is \$10 and the price of labor \$15. What combination of labor and capital minimizes the cost of producing any given output?
- Suppose the process of producing lightweight parkas by Polly's Parkas is described by the function

$$Q = 10K^5(L - 40)^2$$

where Q is the number of parkas produced, K the number of computerized stitching-machine hours, and L the number of person-hours of labor. In addition to capital and labor, \$10 worth of raw materials are used in the production of each parka.

- By minimizing cost subject to the production function, derive the cost-minimizing demands for K and L as a function of output (Q), wage rates (w), and rental rates on machines (r). Use these results to derive the total cost function: that is, costs as a function of Q , r , w , and the constant \$10 per unit materials cost.
- This process requires skilled workers, who earn \$32 per hour. The rental rate on the machines used in the process is \$64 per hour. At these factor prices, what are total costs as a function of Q ? Does this technology exhibit decreasing, constant, or increasing returns to scale?
- Polly's Parkas plans to produce 2000 parkas per week. At the factor prices given above, how many workers should the firm hire (at 40 hours per week) and how many machines should it rent (at 40 machine-hours per week)? What are the marginal and average costs at this level of production?

CHAPTER 8

Profit Maximization and Competitive Supply

A cost curve describes the minimum cost at which a firm can produce various amounts of output. Once we know its cost curve, we can turn to a fundamental problem faced by every firm: How much should be produced? In this chapter, we will see how a perfectly competitive firm chooses the level of output that maximizes its profit. We will also see how the output choices of individual firms lead to a supply curve for an entire industry.

Our discussion of production and cost in Chapters 6 and 7 applies to firms in all kinds of markets. However, in this chapter we focus on firms in *perfectly competitive markets*, in which all firms produce an identical product and each is so small in relation to the industry that its production decisions have no effect on market price. New firms can easily enter the industry if they perceive a potential for profit, and existing firms can exit if they start losing money.

We begin by explaining exactly what is meant by a *competitive market*. We then explain why it makes sense to assume that firms (in any market) have the objective of maximizing profit. We provide a rule for choosing the profit-maximizing output for firms in all markets, competitive or otherwise. Following this we show how a competitive firm chooses its output in the short and long run.

We next examine how the firm's output choice changes as the cost of production or the prices of inputs change. In this way, we show how to derive the *firm's supply curve*. We then aggregate the supply curves of individual firms to obtain the *industry supply curve*. In the short run, firms in an industry choose which level of output to produce to maximize profit. In the long run, they not only make output choices but also decide whether to be in a market at all. We will see that while the prospect of high profits encourages firms to enter an industry, losses encourage them to leave.

Chapter Outline

- Perfectly Competitive Markets 252
- Profit Maximization 254
- Marginal Revenue, Marginal Cost, and Profit Maximization 255
- Choosing Output in the Short Run 258
- The Competitive Firm's Short-Run Supply Curve 263
- The Short-Run Market Supply Curve 266
- Choosing Output in the Long Run 271
- The Industry's Long-Run Supply Curve 277

List of Examples

- The Short-Run Output Decision of an Aluminum Smelting Plant 260
- Some Cost Considerations for Managers 261
- The Short-Run Production of Petroleum Products 265
- The Short-Run World Supply of Copper 268
- The Long-Run Supply of Housing 282

8.1 Perfectly Competitive Markets

In Chapter 2, we used supply-demand analysis to explain how changing market conditions affect the market price of such products as wheat and gasoline. We saw that the equilibrium price and quantity of each product was determined by the intersection of the market demand and market supply curves. Underlying this analysis is the model of a *perfectly competitive market*. The model of perfect competition is very useful for studying a variety of markets, including agriculture, fuels and other commodities, housing, services, and financial markets. Because this model is so important, we will spend some time laying out the basic assumptions that underlie it.

The model of perfect competition rests on three basic assumptions: (1) price taking, (2) product homogeneity, and (3) free entry and exit. You have encountered these assumptions earlier in the book; here we summarize and elaborate on them.

Price Taking Many firms compete in the market, and therefore each firm faces a significant number of direct competitors for its products. Because *each individual firm sells a sufficiently small proportion of total market output, its decisions have no impact on market price*. Thus each firm *takes the market price as given*. In short, firms in perfectly competitive markets are **price takers**.

price taker Firm that has no influence over market price and that thus takes the price as a given.

Suppose, for example, that you are the owner of a small electric lightbulb distribution business. You buy your lightbulbs from the manufacturer and resell them at wholesale to small businesses and retail outlets. Unfortunately, you are only one of many competing distributors. As a result, you find that there is little room to negotiate with your customers. If you do not offer a competitive price—one that is determined in the marketplace—your customers will take their business elsewhere. In addition, you know that the number of lightbulbs that you sell will have little or no effect on the wholesale price of bulbs. You are a price taker.

The assumption of price taking applies to *consumers* as well as firms. In a perfectly competitive market, each consumer buys such a small proportion of total industry output that he or she has no impact on the market price, and therefore takes the price as given.

Another way of stating the price-taking assumption is that there are many independent firms and independent consumers in the market, all of whom believe—correctly—that their decisions will not affect prices.

Product Homogeneity Price-taking behavior typically occurs in markets where firms produce identical, or nearly identical, products. When *the products of all of the firms in a market are perfectly substitutable with one another*—that is, when they are *homogeneous*—no firm can raise the price of its product above the price of other firms without losing most or all of its business. Most agricultural products are homogeneous: Because product quality is relatively similar among farms in a given region, for example, buyers of corn do not ask which individual farm grew the product. Oil, gasoline, and raw materials such as copper, iron, lumber, cotton, and sheet steel are also fairly homogeneous. Economists refer to such homogeneous products as *commodities*.

In contrast, when products are not homogeneous, each firm has the opportunity to raise its price above that of its competitors without losing all of its sales. Premium ice creams such as Haagen-Daaz, for example, can be sold at higher prices because Haagen-Daaz has different ingredients and is perceived by many consumers as a higher-quality product.

The assumption of product homogeneity is important because it ensures that there is a *single market price*, consistent with supply-demand analysis.

Free Entry and Exit This third assumption, of **free entry (exit)**, means that there are no special costs that make it difficult for a new firm either to enter an industry and produce or to exit if it cannot make a profit. *As a result, buyers can easily switch from one supplier to another, and suppliers can easily enter or exit a market.*

free entry (exit) When there are no special costs that make it difficult for a firm to enter (or exit) an industry.

The special costs that could restrict entry are costs that an entrant to a market would have to bear but a firm that is already producing will not. The pharmaceutical industry, for example, is not perfectly competitive because Merck, Pfizer, and other firms hold patents that give them unique rights to produce drugs. Any new entrant would either have to invest in research and development to obtain its own competing drugs or pay substantial license fees to one or more firms already in the market. R&D expenditures or license fees could limit a firm's ability to enter the market. Likewise, the aircraft industry is not perfectly competitive because entry requires an immense investment in plant and equipment that has little or no resale value.

The assumption of free entry and exit is important for competition to be effective. It means that consumers can easily switch to a rival firm if a current supplier raises its price. For businesses, it means that a firm can freely enter an industry if it sees a profit opportunity and exit if it is losing money. Thus a firm can hire labor and purchase capital and raw materials as needed, and it can release or relocate these factors of production if it wants to shut down or relocate.

If these three assumptions of perfect competition hold, market demand and supply curves can be used to analyze the behavior of market prices. In most markets, of course, these assumptions are unlikely to hold exactly. This does not mean, however, that the model of perfect competition is not useful. Some markets do indeed come close to satisfying our assumptions. But even when one or more of these three assumptions fails to hold, so that a market is not perfectly competitive, much can be learned by making comparisons with the perfectly competitive ideal.

When Is a Market Highly Competitive?

Apart from agriculture, few real-world markets are *perfectly* competitive in the sense that each firm faces a perfectly horizontal demand curve for a homogeneous product in an industry that it can freely enter or exit. Nevertheless, many markets are *highly* competitive in the sense that firms face highly elastic demand curves and relatively easy entry and exit.

A simple rule of thumb to describe whether a market is close to being perfectly competitive would be appealing. Unfortunately, we have no such rule, and it is important to understand why. Consider the most obvious candidate: an industry with many firms (say, at least 10 to 20). Because firms can implicitly or explicitly collude in setting prices, the presence of many firms is not sufficient for an industry to approximate perfect competition. Conversely, the presence of only a few firms in a market does not rule out competitive behavior. Suppose that only three firms are in the market but that market demand for the product is very elastic. In this case, the demand curve facing each firm is likely to be nearly horizontal and the firms will behave *as if* they were operating in a perfectly competitive market. Even if market demand is not very elastic, these three firms

might compete very aggressively (as we will see in Chapter 13). The important point to remember is that although firms may behave competitively in many situations, there is no simple indicator to tell us when a market is highly competitive. Often it is necessary to analyze both the firms themselves and their strategic interactions, as we do in Chapters 12 and 13.

8.2 Profit Maximization

We now turn to the analysis of profit maximization. In this section, we ask whether firms do indeed seek to maximize profit. Then in Section 8.3 we will describe a rule that any firm—whether in a competitive market or not—can use to find its profit-maximizing output level. Then, we will consider the special case of a firm in a competitive market. We distinguish the demand curve facing a competitive firm from the market demand curve and use this information to describe the competitive firm's profit-maximization rule.

Do Firms Maximize Profit?

The assumption of *profit maximization* is frequently used in microeconomics because it predicts business behavior reasonably accurately and avoids unnecessary analytical complications. But the question of whether firms actually do seek to maximize profit has been controversial.

For smaller firms managed by their owners, profit is likely to dominate almost all decisions. In larger firms, however, managers who make day-to-day decisions usually have little contact with the owners (i.e., the stockholders). As a result, owners cannot monitor the managers' behavior on a regular basis. Managers then have some leeway in how they run the firm and can deviate from profit-maximizing behavior.

Managers may be more concerned with such goals as revenue maximization, revenue growth, or the payment of dividends to satisfy shareholders. They might also be overly concerned with the firm's short-run profit (perhaps to earn a promotion or a large bonus) at the expense of its longer-run profit, even though long-run profit maximization better serves the interests of the stockholders.¹ (We discuss the implications of differences between the incentives of managers and owners in greater detail in Chapter 17.)

Even so, managers' freedom to pursue goals other than long-run profit maximization is limited. If they do pursue such goals, shareholders or boards of directors can replace them, or the firm can be taken over by new management. In any case, firms that do not come close to maximizing profit are not likely to survive. Firms that do survive in competitive industries make long-run profit maximization one of their highest priorities.

Thus our working assumption of profit maximization is reasonable. Firms that have been in business for a long time are likely to care a lot about profit, whatever else their managers may appear to be doing. For example, a firm that

¹ To be more exact, maximizing the market value of the firm is a more appropriate goal than profit maximization because market value includes the stream of profits that the firm earns over time. It is the stream of current and future profits that is of direct interest to the stockholders.

subsidizes public television may seem public-spirited and altruistic. Yet this beneficence is likely to be in the long-run financial interest of the firm because it generates goodwill for the firm and its products.

8.3 Marginal Revenue, Marginal Cost, and Profit Maximization

Let's begin by looking at the profit-maximizing output decision for *any* firm, whether the firm operates in a perfectly competitive market or is one that can influence price. Because **profit** is the difference between (total) revenue and (total) cost, finding the firm's profit-maximizing output level means analyzing its revenue. Suppose that the firm's output is q , and that it obtains revenue R . This revenue is equal to the price of the product P times the number of units sold: $R = Pq$. The cost of production C also depends on the level of output. The firm's profit, π , is the difference between revenue and cost:

$$\pi(q) = R(q) - C(q)$$

(Here we show explicitly that π , R , and C depend on output. Usually we will omit this reminder.)

To maximize profit, the firm selects the output for which the difference between revenue and cost is the greatest. This principle is illustrated in Figure 8.1. Revenue $R(q)$ is a curved line, which reflects the fact that the firm can sell a greater level of output only by lowering its price. The slope of this revenue curve

profit Difference between total revenue and total cost.

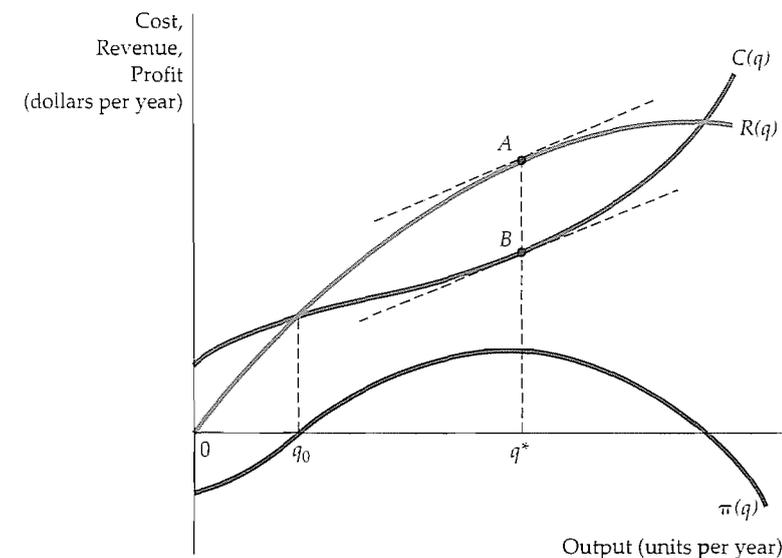


FIGURE 8.1 Profit Maximization in the Short Run

A firm chooses output q^* , so that profit, the difference AB between revenue R and cost C , is maximized. At that output, marginal revenue (the slope of the revenue curve) is equal to marginal cost (the slope of the cost curve).

marginal revenue Change in revenue resulting from a one-unit increase in output.

is **marginal revenue**: the change in revenue resulting from a one-unit increase in output.

Also shown is the total cost curve $C(q)$. The slope of this curve, which measures the additional cost of producing one additional unit of output, is the firm's *marginal cost*. Note that total cost $C(q)$ is positive when output is zero because there is a fixed cost in the short run.

For the firm illustrated in Figure 8.1, profit is negative at low levels of output, because revenue is insufficient to cover fixed and variable costs. As output increases, revenue rises more rapidly than cost, so that profit eventually becomes positive. Profit continues to increase until output reaches the level q^* . At this point, marginal revenue and marginal cost are equal, and the vertical distance between revenue and cost, AB , is greatest. q^* is the profit-maximizing output level. Note that at output levels above q^* , cost rises more rapidly than revenue—i.e., marginal revenue is less than marginal cost. Thus, profit declines from its maximum when output increases above q^* .

The rule that profit is maximized when marginal revenue is equal to marginal cost holds for all firms, whether competitive or not. This important rule can also be derived algebraically. Profit, $\pi = R - C$, is maximized at the point at which an additional increment to output leaves profit unchanged (i.e., $\Delta\pi/\Delta q = 0$):

$$\Delta\pi/\Delta q = \Delta R/\Delta q - \Delta C/\Delta q = 0$$

$\Delta R/\Delta q$ is marginal revenue MR and $\Delta C/\Delta q$ is marginal cost MC. Thus we conclude that profit is maximized when $MR - MC = 0$, so that

$$MR(q) = MC(q)$$

Demand and Marginal Revenue for a Competitive Firm

Because each firm in a competitive industry sells only a small fraction of the entire industry sales, *how much output the firm decides to sell will have no effect on the market price of the product*. The market price is determined by the industry demand and supply curves. Therefore, the competitive firm is a *price taker*. Recall that price taking is one of the fundamental assumptions of perfect competition. The price-taking firm knows that its production decision will have no effect on the price of the product. For example, when a farmer is deciding how many acres of wheat to plant in a given year, he can take the market price of wheat—say, \$4 per bushel—as given. That price will not be affected by his acreage decision.

Often we will want to distinguish between market demand curves and the demand curves that individual firms face. In this chapter we will denote *market* output and demand by capital letters (Q and D), and the *firm's* output and demand by lowercase letters (q and d).

Because it is a price taker, *the demand curve d facing an individual competitive firm is given by a horizontal line*. In Figure 8.2(a), the farmer's demand curve corresponds to a price of \$4 per bushel of wheat. The horizontal axis measures the amount of wheat that the farmer can sell, and the vertical axis measures the price.

Compare the demand curve facing the firm (in this case the farmer) in Figure 8.2(a) with the market demand curve D in Figure 8.2(b). The market demand curve shows how much wheat *all consumers* will buy at each possible price. It is downward sloping because consumers buy more wheat at a lower price. The demand curve facing the firm, however, is horizontal because the firm's sales will have

In §4.1, we explain how the demand curve relates the quantity of a good that a consumer will buy to the price of that good.

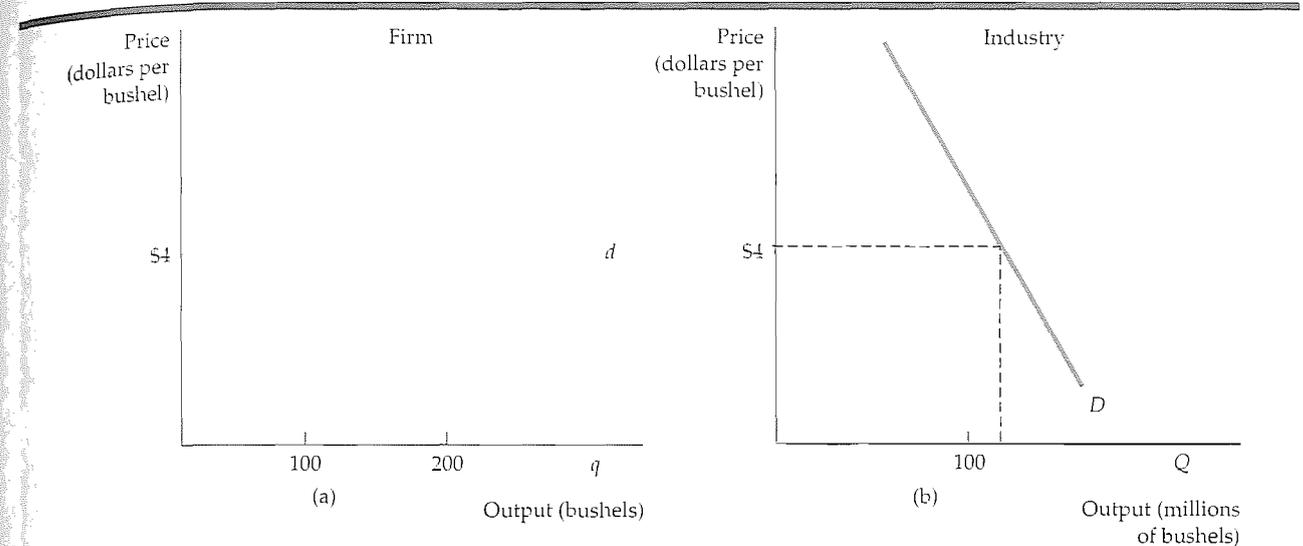


FIGURE 8.2 Demand Curve Faced by a Competitive Firm

A competitive firm supplies only a small portion of the total output of all the firms in an industry. Therefore the firm takes the market price of the product as given, choosing its output on the assumption that the price will be unaffected by the output choice. In (a) the demand curve facing the firm is perfectly elastic, even though the market demand curve in (b) is downward sloping.

no effect on price. Suppose the firm increased its sales from 100 to 200 bushels of wheat. This would have almost no effect on the market because the industry output of wheat is 100 million bushels. Price is determined by the interaction of all firms and consumers in the market, not by the output decision of a single firm.

By the same token, when an individual firm faces a horizontal demand curve, it can sell an additional unit of output without lowering price. As a result, when it sells an additional unit, the firm's *total revenue* increases by an amount equal to the price: one bushel of wheat sold for \$4 yields additional revenue of \$4. Thus, marginal revenue is constant at \$4. At the same time, *average revenue* received by the firm is also \$4 because every bushel of wheat produced will be sold at \$4. Therefore:

The demand curve d facing an individual firm in a competitive market is both its average revenue curve and its marginal revenue curve. Along this demand curve, marginal revenue, average revenue, and price are all equal.

Profit Maximization by a Competitive Firm

Because the demand curve facing a competitive firm is horizontal, so that $MR = P$, the general rule for profit maximization that applies to any firm can be simplified. A perfectly competitive firm should choose its output so that *marginal cost equals price*:

$$MC(q) = MR = P$$

Note that because competitive firms take price as fixed, this is a rule for setting output, not price.

The choice of the profit-maximizing output by a competitive firm is so important that we will devote most of the rest of this chapter to analyzing it. We begin with the short-run output decision and then move to the long run.

8.4 Choosing Output in the Short Run

How much output should a firm produce over the short run, when the firm's plant size is fixed? In this section we show how a firm can use information about revenue and cost to make a profit-maximizing output decision.

Short-Run Profit Maximization by a Competitive Firm

In the short run, a firm operates with a fixed amount of capital and must choose the levels of its variable inputs (labor and materials) to maximize profit. Figure 8.3 shows the firm's short-run decision. The average and marginal revenue curves are drawn as a horizontal line at a price equal to \$40. In this figure, we have drawn the average total cost curve ATC, the average variable cost curve AVC, and the marginal cost curve MC, so that we can see the firm's profit more easily.

Marginal, average, and total cost are discussed in §7.2.

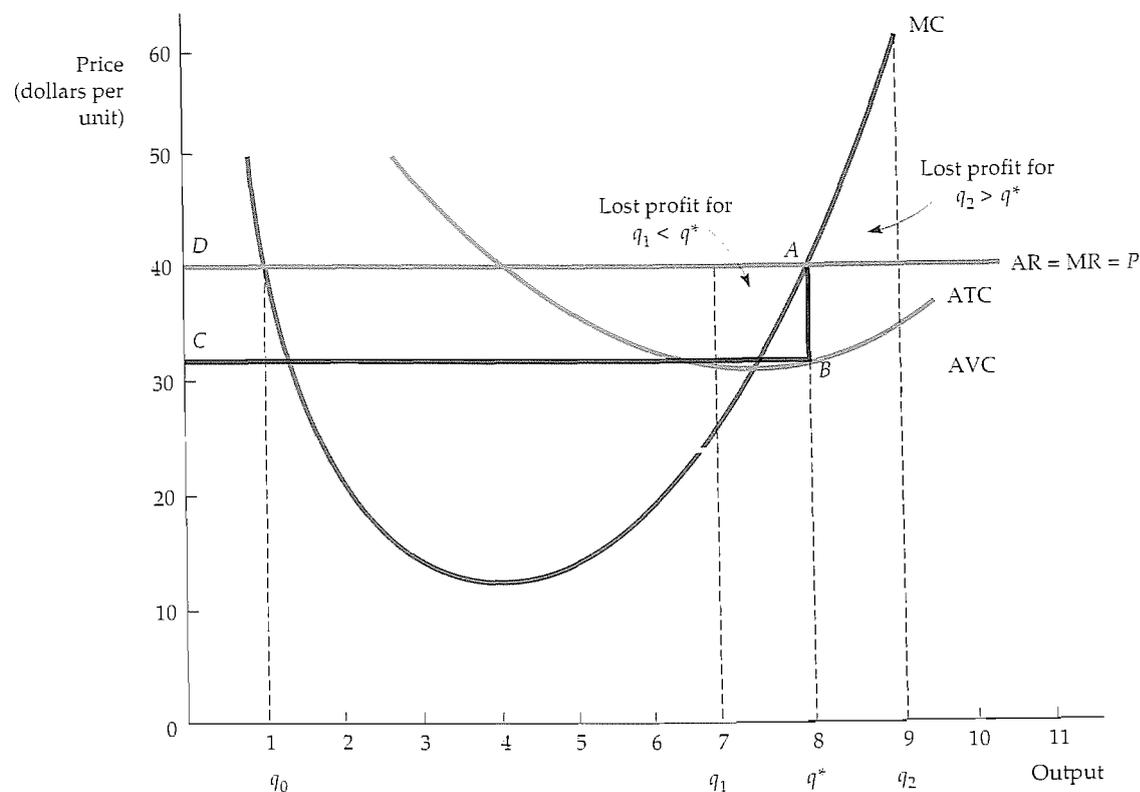


FIGURE 8.3 A Competitive Firm Making a Positive Profit

In the short run, the competitive firm maximizes its profit by choosing an output q^* at which its marginal cost MC is equal to the price P (or marginal revenue MR) of its product. The profit of the firm is measured by the rectangle $ABCD$. Any lower output q_1 , or higher output q_2 , will lead to lower profit.

Profit is maximized at point A , where output is $q^* = 8$ and the price is \$40, because marginal revenue is equal to marginal cost at this point. To see that $q^* = 8$ is indeed the profit-maximizing output, note that at a lower output, say $q_1 = 7$, marginal revenue is greater than marginal cost; profit could thus be increased by increasing output. The shaded area between $q_1 = 7$ and q^* shows the lost profit associated with producing at q_1 . At a higher output, say q_2 , marginal cost is greater than marginal revenue; thus, reducing output saves a cost that exceeds the reduction in revenue. The shaded area between q^* and $q_2 = 9$ shows the lost profit associated with producing at q_2 .

The MR and MC curves cross at an output of q_0 as well as q^* . At q_0 , however, profit is clearly not maximized. An increase in output beyond q_0 increases profit because marginal cost is well below marginal revenue. We can thus state the condition for profit maximization as follows: *Marginal revenue equals marginal cost at a point at which the marginal cost curve is rising.* This conclusion is very important because it applies to the output decisions of firms in markets that may or may not be perfectly competitive. We can restate it as follows:

Output Rule: If a firm is producing any output at all, it should produce at the level at which marginal revenue equals marginal cost.

The Short-Run Profit of a Competitive Firm

Figure 8.3 also shows the competitive firm's short-run profit. The distance AB is the difference between price and average cost at the output level q^* , which is the average profit per unit of output. Segment BC measures the total number of units produced. Rectangle $ABCD$, therefore, is the firm's profit.

A firm need not always earn a profit in the short run, as Figure 8.4 shows. The major difference from Figure 8.3 is a higher fixed cost of production. This higher fixed cost raises average total cost but does not change the average variable cost and marginal cost curves. At the profit-maximizing output q^* , the price P is less than average cost. Line segment AB , therefore, measures the average loss from production. Likewise, the rectangle $ABCD$ now measures the firm's total loss.

Why doesn't a firm that earns a loss leave an industry entirely? A firm might operate at a loss in the short run because it expects to earn a profit in the future, when the price of its product increases or the cost of production falls, and because shutting down and starting up again would be costly. In fact, a firm has two choices in the short run: It can produce some output, or it can shut down production temporarily. It will compare the profitability of producing with the profitability of shutting down and choose the preferred outcome. *If the price of the product is greater than the average economic cost of production, the firm makes a positive economic profit by producing. Consequently, it will choose to produce.*

But suppose that the price is less than average total cost, as shown in Figure 8.4. If it continues to produce, the firm minimizes its losses at output q^* . Note that in Figure 8.4, because of the presence of fixed costs, average variable cost is less than average total cost and the firm is indeed losing money. The firm should therefore consider shutting down. If it does, it earns no revenue, but it avoids the fixed as well as variable cost of production. If there are no sunk costs so that average economic cost is equal to average total cost, the firm should indeed shut down. Because there are no sunk costs, it can invest its capital elsewhere or, for that matter, reenter the industry if and when economic conditions improve.

To summarize: When there are no sunk costs, the firm's average total cost is equal to its average economic cost. Thus, *the firm should shut down when the price of its product is less than the average total cost at the profit-maximizing output.*

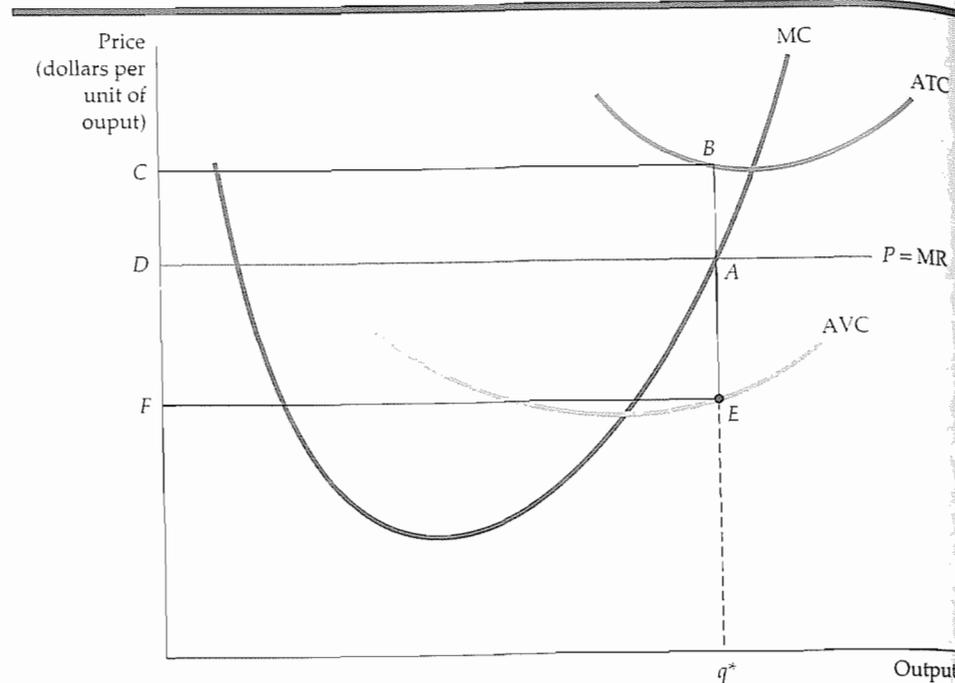


FIGURE 8.4 A Competitive Firm Incurring Losses

A competitive firm should shut down if price is below ATC. If the firm has sunk costs that it amortizes and treats as fixed, it may produce in the short run if price is greater than average variable cost.

Suppose, instead, that the firm has paid a large sunk cost, which it amortizes and treats as an ongoing fixed cost. In this case, the rectangle CBEF in Figure 8.4 represents a component of total cost that cannot be avoided even if the firm shuts down. (The firm's capital investment will be of no value if it shuts down.) As a result, the firm's average variable cost is now the appropriate measure of the firm's average economic cost of production. Therefore, *the firm should stay in business as long as the price of its product is greater than its average variable cost of production at the profit-maximizing output.*

Note that whether or not the firm has sunk costs, there is one shut-down rule that always applies:

Shut-Down Rule: The firm should shut down if the price of the product is less than the average economic cost of production at the profit-maximizing output.

EXAMPLE 8.1 The Short-Run Output Decision of an Aluminum Smelting Plant

How should the manager of an aluminum smelting plant determine the plant's profit-maximizing output? Recall from Example 7.3 that the smelting plant's short-run marginal cost of production depends on whether it is running two or three shifts per day. As shown in Figure 8.5, marginal cost is \$1140 per ton for output levels up to 600 tons per day and \$1300 per ton for output levels between 600 and 900 tons per day.

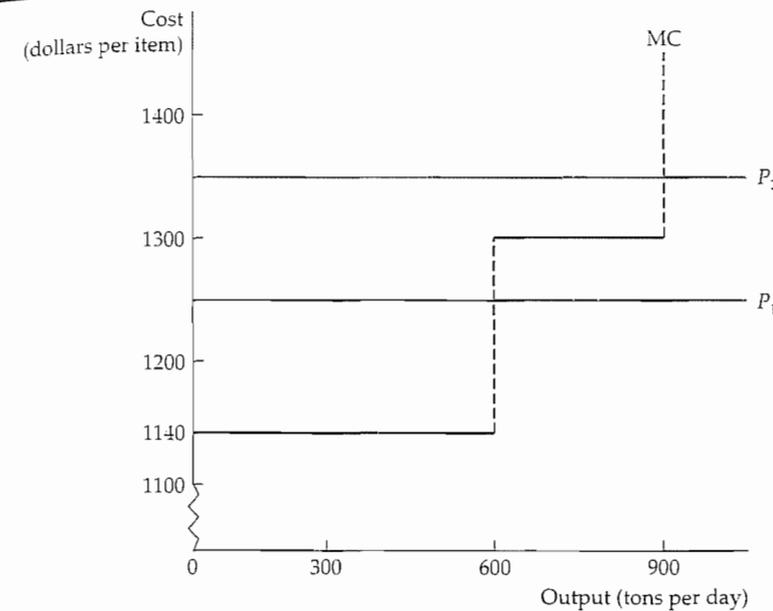


FIGURE 8.5 The Short-Run Output of an Aluminum Smelting Plant

In the short run, the plant should produce 600 tons per day if price is above \$1140 per ton but less than \$1300 per ton. If price is greater than \$1300 per ton, it should run an overtime shift and produce 900 tons per day. If price drops below \$1140 per ton, the firm should stop producing, but it should probably stay in business because the price may rise in the future.

Suppose that the price of aluminum is initially $P_1 = \$1250$ per ton. In that case, the profit-maximizing output is 600 tons; the firm can make a profit above its variable cost of \$110 per ton by employing workers for two shifts a day. Running a third shift would involve overtime, and the price of the aluminum is insufficient to make the added production profitable. Suppose, however, that the price of aluminum were to increase to $P_2 = \$1360$ per ton. This price is greater than the \$1300 marginal cost of the third shift, making it profitable to increase output to 900 tons per day.

Finally, suppose the price drops to only \$1100 per ton. In this case, the firm should stop producing, but it should probably stay in business. By taking this step, it could resume producing in the future should the price increase.

EXAMPLE 8.2 Some Cost Considerations for Managers

The application of the rule that marginal revenue should equal marginal cost depends on a manager's ability to estimate marginal cost.² To obtain useful measures of cost, managers should keep three guidelines in mind.

² This example draws on the discussion of costs and managerial decision making in Thomas Nagle and Reed Holden, *The Strategy and Tactics of Pricing*, 2nd ed. (Englewood Cliffs, NJ: Prentice-Hall, 1995), ch. 2.

First, except under limited circumstances, *average variable cost should not be used as a substitute for marginal cost*. When marginal and average cost are nearly constant, there is little difference between them. However, if both marginal and average cost are increasing sharply, the use of average variable cost can be misleading in deciding how much to produce. Suppose for example, that a company has the following cost information:

Current output	100 units per day, 80 of which are produced during the regular shift and 20 of which are produced during overtime
Materials cost	\$8 per unit for all output
Labor cost	\$30 per unit for the regular shift; \$50 per unit for the overtime shift

Let's calculate average variable cost and marginal cost for the first 80 units of output and then see how both cost measures change when we include the additional 20 units produced with overtime labor. For the first 80 units, average variable cost is simply the labor cost ($\$2400 = \$30 \text{ per unit} \times 80 \text{ units}$) plus the materials cost ($\$640 = \$8 \text{ per unit} \times 80 \text{ units}$) divided by the 80 units— $(\$2400 + \$640)/80 = \$38 \text{ per unit}$. Because the average variable cost is the same for each unit of output, the marginal cost is also equal to \$38 per unit.

When output increases to 100 units per day, both average variable cost and marginal cost change. The variable cost has now increased; it includes the additional materials cost of \$160 (20 units \times \$8 per unit) and the additional labor cost of \$1000 (20 units \times \$50 per unit). Average variable cost is therefore the total labor cost plus the materials cost ($\$2400 + \$1000 + \$640 + \160) divided by the 100 units of output, or \$42 per unit.

What about marginal cost? While the materials cost per unit has remained unchanged at \$8 per unit, the marginal cost of labor has now increased to \$50 per unit, so that the marginal cost of each unit of overtime output is \$58 per day. Because the marginal cost is higher than the average variable cost, a manager who relies on average variable cost will produce too much output.

Second, *a single item on a firm's accounting ledger may have two components, only one of which involves marginal cost*. Suppose, for example, that a manager is trying to cut back production. She reduces the number of hours that some employees work and lays off others. But the salary of an employee who is laid off may not be an accurate measure of the marginal cost of production when cuts are made. Union contracts, for example, often require the firm to pay laid-off employees part of their salary. In this case, the marginal cost of increasing production is not the same as the savings in marginal cost when production is decreased. The savings is the labor cost after the required layoff salary has been subtracted.

Third, *all opportunity costs should be included in determining marginal cost*. Suppose a department store wants to sell children's furniture. Instead of building a new selling area, the manager decides to use part of the third floor, which had been used for appliances, for the furniture. The marginal cost of this space is the \$90 per square foot per day in profit that would have been earned had the store continued to sell appliances there. This opportunity cost measure may be much greater than what the store actually paid for that part of the building.

These three guidelines can help a manager to measure marginal cost correctly. Failure to do so can cause production to be too high or too low and thereby reduce profit.

8.5 The Competitive Firm's Short-Run Supply Curve

A *supply curve* for a firm tells us how much output it will produce at every possible price. We have seen that competitive firms will increase output to the point at which price is equal to marginal cost but will shut down if price is below average economic cost. We have also seen that average economic cost is equal to average total cost when there are no sunk costs but equal to average variable cost when costs treated as fixed are actually amortized sunk costs. Therefore, the firm's supply curve is *the portion of the marginal cost curve that lies above the average economic cost curve*.

Figure 8.6 illustrates the short-run supply curve for the case in which all fixed costs are actually amortized sunk costs. In this case, for any P greater than minimum AVC, the profit-maximizing output can be read directly from the graph. At

In §7.1, we explain that economic cost is the cost associated with forgone opportunities.

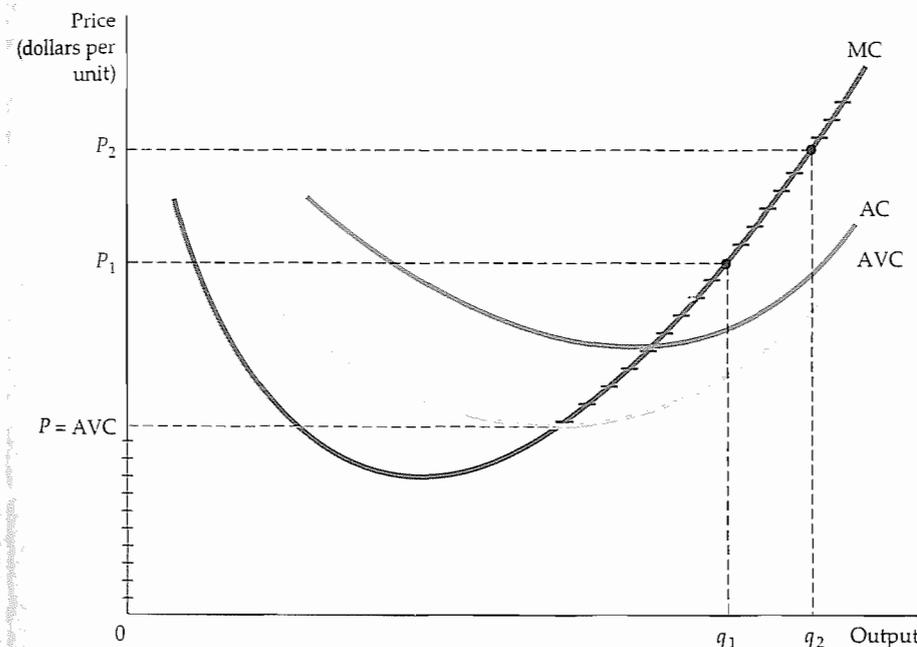


FIGURE 8.6 The Short-Run Supply Curve for a Competitive Firm

In the short run, the firm chooses its output so that marginal cost MC is equal to price as long as the firm covers its average economic cost. When all fixed costs are amortized sunk costs, the short-run supply curve is given by the crosshatched portion of the marginal cost curve.

a price P_1 , for example, the quantity supplied will be q_1 ; and at P_2 , it will be q_2 . For P less than (or equal to) minimum AVC, the profit-maximizing output is equal to zero. In Figure 8.6 the entire short-run supply curve consists of the crosshatched portion of the vertical axis plus the marginal cost curve above the point of minimum average variable cost.

Short-run supply curves for competitive firms slope upward for the same reason that marginal cost increases—the presence of diminishing marginal returns to one or more factors of production. As a result, an increase in the market price will induce those firms already in the market to increase the quantities they produce. The higher price makes the additional production profitable and also increases the firm's total profit because it applies to all units that the firm produces.

In §6.3, we explain that diminishing marginal returns occurs when each additional increase in an input results in a smaller and smaller increase in output.

The Firm's Response to an Input Price Change

When the price of its product changes, the firm changes its output level to ensure that marginal cost of production remains equal to price. Often, however, the product price changes at the same time that the prices of inputs change. In this section we show how the firm's output decision changes in response to a change in the price of one of its inputs.

Figure 8.7 shows a firm's marginal cost curve that is initially given by MC_1 when the firm faces a price of \$5 for its product. The firm maximizes profit by producing an output of q_1 . Now suppose the price of one input increases. Because it now costs more to produce each unit of output, this increase causes the marginal cost curve to shift upward from MC_1 to MC_2 . The new profit-maximizing output is q_2 , at which $P = MC_2$. Thus, the higher input price causes the firm to reduce its output.

If the firm had continued to produce q_1 , it would have incurred a loss on the last unit of production. In fact, all production beyond q_2 reduces profit. The shaded area in the figure gives the total savings to the firm (or equivalently, the reduction in lost profit) associated with the reduction in output from q_1 to q_2 .

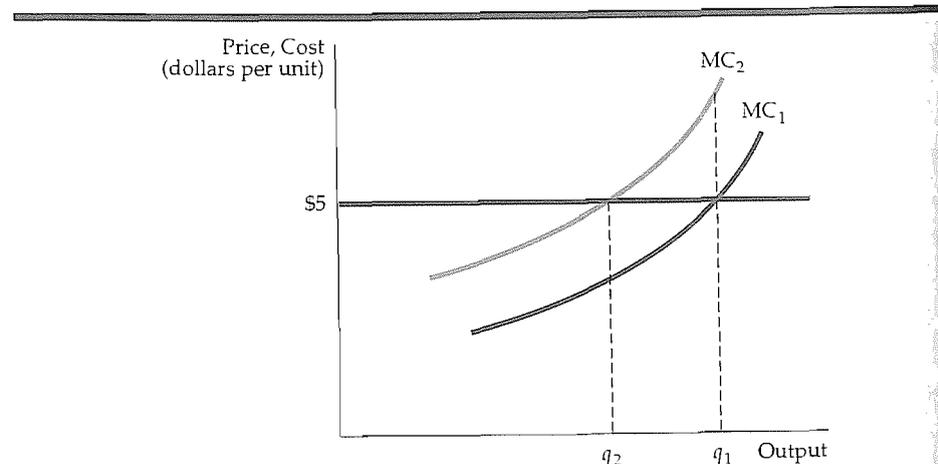


FIGURE 8.7 The Response of a Firm to a Change in Input Price

When the marginal cost of production for a firm increases (from MC_1 to MC_2), the level of output that maximizes profit falls (from q_1 to q_2).

EXAMPLE 8.3 The Short-Run Production of Petroleum Products

Suppose you are managing an oil refinery that converts crude oil into a particular mix of products, including gasoline, jet fuel, and residual fuel oil for home heating. Although plenty of crude oil is available, the amount that you refine depends on the capacity of the refinery and the cost of production. How much should you produce each day?³

Information about the refinery's marginal cost of production is essential for this decision. Figure 8.8 shows the short-run marginal cost curve (SMC). Marginal cost increases with output, but in a series of uneven segments rather than as a smooth curve. The increase occurs in segments because the refinery uses different processing units to turn crude oil into finished products. When a particular processing unit reaches capacity, output can be increased only by substituting a more expensive process. For example, gasoline can be produced from light crude oils rather inexpensively in a processing unit called a "thermal cracker." When this unit becomes full, additional gasoline can still be produced (from heavy as well as light crude oil), but only at a higher cost. In the case

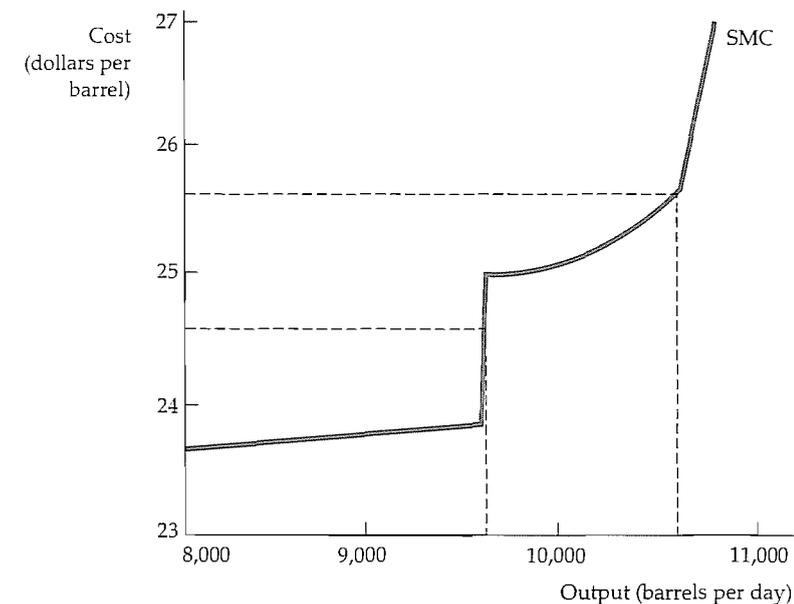


FIGURE 8.8 The Short-Run Production of Petroleum Products

The marginal cost of producing petroleum products from crude oil increases sharply at several levels of output as the refinery shifts from one processing unit to another. As a result, the output level can be insensitive to some changes in price but very sensitive to others.

³ This example is based on James M. Griffin, "The Process Analysis Alternative to Statistical Cost Functions: An Application to Petroleum Refining," *American Economic Review* 62 (1972): 46–56. The numbers have been updated and applied to a particular refinery.

illustrated by Figure 8.8 the first capacity constraint comes into effect when production reaches about 9700 barrels a day. A second capacity constraint becomes important when production increases beyond 10,700 barrels a day.

Deciding how much output to produce now becomes relatively easy. Suppose that refined products can be sold for \$23 per barrel. Because the marginal cost of production is close to \$24 for the first unit of output, no crude oil should be run through the refinery at a price of \$23. If, however, price is between \$24 and \$25, the refinery should produce 9700 barrels a day (filling the thermal cracker). Finally, if the price is above \$25, the more expensive refining unit should be used and production expanded toward 10,700 barrels a day.

Because the cost function rises in steps, you know that your production decisions need not change much in response to *small* changes in price. You will typically utilize sufficient crude oil to fill the appropriate processing unit until price increases (or decreases) substantially. In that case, you need simply calculate whether the increased price warrants using an additional, more expensive processing unit.

8.6 The Short-Run Market Supply Curve

The *short-run market supply curve* shows the amount of output that the industry will produce in the short run for every possible price. The industry's output is the sum of the quantities supplied by all of its individual firms. Therefore, the market supply curve can be obtained by adding the supply curves of each of these firms. Figure 8.9 shows how this is done when there are only three firms, all of which have different short-run production costs. Each firm's marginal cost curve is drawn only for the portion that lies above its average variable cost curve. (We have shown only three firms to keep the graph simple, but the same analysis applies when there are many firms.)

At any price below P_1 , the industry will produce no output because P_1 is the minimum average variable cost of the lowest-cost firm. Between P_1 and P_2 , only firm 3 will produce. The industry supply curve, therefore, will be identical to that portion of firm 3's marginal cost curve MC_3 . At price P_2 , the industry supply will be the sum of the quantity supplied by all three firms. Firm 1 supplies 2 units, firm 2 supplies 5 units, and firm 3 supplies 8 units. Industry supply is thus 15 units. At price P_3 , firm 1 supplies 4 units, firm 2 supplies 7 units, and firm 3 supplies 10 units; the industry supplies 21 units. Note that the industry supply curve is upward sloping but has a kink at price P_2 , the lowest price at which all three firms produce. With many firms in the market, however, the kink becomes unimportant. Thus we usually draw industry supply as a smooth, upward-sloping curve.

Elasticity of Market Supply

Unfortunately, finding the industry supply curve is not always as simple as adding up a set of individual supply curves. As price rises, all firms in the industry expand their output. This additional output increases the demand for inputs

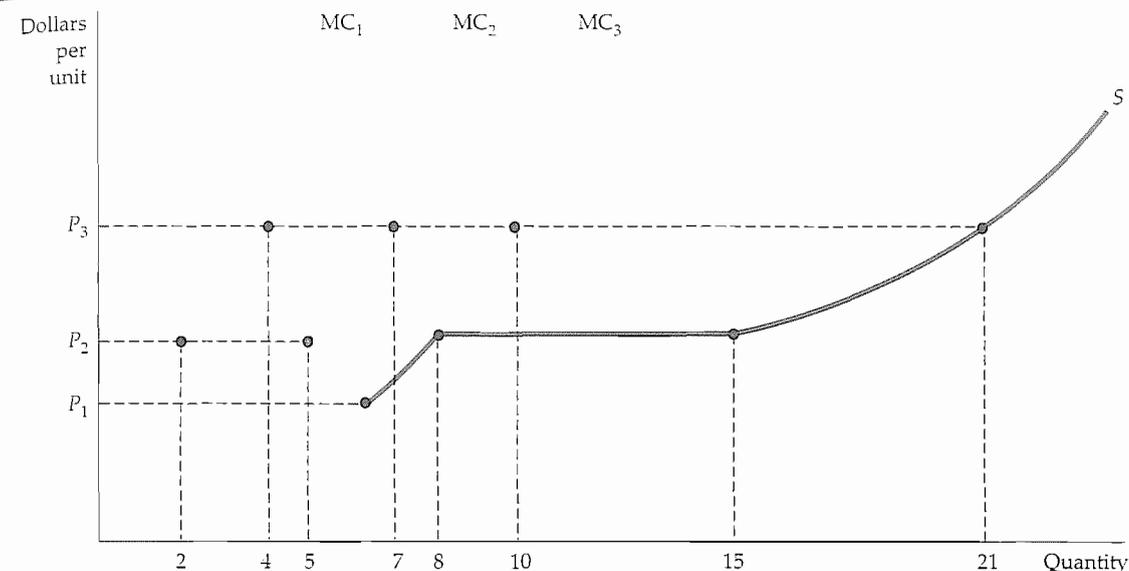


FIGURE 8.9 Industry Supply in the Short Run

The short-run industry supply curve is the summation of the supply curves of the individual firms. Because the third firm has a lower average variable cost curve than the first two firms, the market supply curve S begins at price P_1 and follows the marginal cost curve of the third firm MC_3 until price equals P_2 , where there is a kink. For P_2 and all prices above it, the industry quantity supplied is the sum of the quantities supplied by each of the three firms.

to production and may lead to higher input prices. As we saw in Figure 8.7, increasing input prices shifts the firms' marginal cost curves upward. For example, an increased demand for beef could also increase demand for corn and soybeans (which are used to feed cattle) and thereby cause the prices of these crops to rise. In turn, the higher input prices would cause firms' marginal cost curves to shift upward. This increase lowers each firm's output choice (for any given market price) and causes the industry supply curve to be less responsive to changes in output price than it would otherwise be.

The price elasticity of market supply measures the sensitivity of industry output to market price. The elasticity of supply E_s is the percentage change in quantity supplied Q in response to a 1-percent change in price P :

$$E_s = (\Delta Q/Q)/(\Delta P/P)$$

Because marginal cost curves are upward sloping, the short-run elasticity of supply is always positive. When marginal costs increase rapidly in response to increases in output, the elasticity of supply is low. Firms are then capacity-constrained and find it costly to increase output. But when marginal costs increase slowly in response to increases in output, supply is relatively elastic; in this case, a small price increase induces firms to produce much more.

At one extreme is the case of *perfectly inelastic supply*, which arises when the industry's plant and equipment are so fully utilized that greater output can be achieved only if new plants are built (as they will be in the long run). At the other extreme is the case of *perfectly elastic supply*, which arises when marginal costs are constant.

In §2.3, we define the elasticity of supply as the percentage change in quantity supplied resulting from a 1-percent increase in price.

EXAMPLE 8.4 The Short-Run World Supply of Copper

In the short run, the shape of the market supply curve for a mineral such as copper depends on how the cost of mining varies within and among the world's major producers. Costs of mining, smelting, and refining copper differ because of differences in labor and transportation costs and because of differences in the copper content of the ore. Table 8.1 summarizes some of the relevant cost and production data for the nine largest copper-producing nations.⁴

These data can be used to plot the world supply curve for copper. The supply curve is a short-run curve because it takes the existing mines and refineries as fixed. Figure 8.10 shows how this curve is constructed for the nine countries listed in the table. The complete world supply curve would, of course, incorporate data for all copper-producing countries. Note also that the curve in Figure 8.10 is an approximation. The marginal cost number for each country is an average for all copper producers in that country. In the United States, for example, some producers have a marginal cost greater than 70 cents and some less.

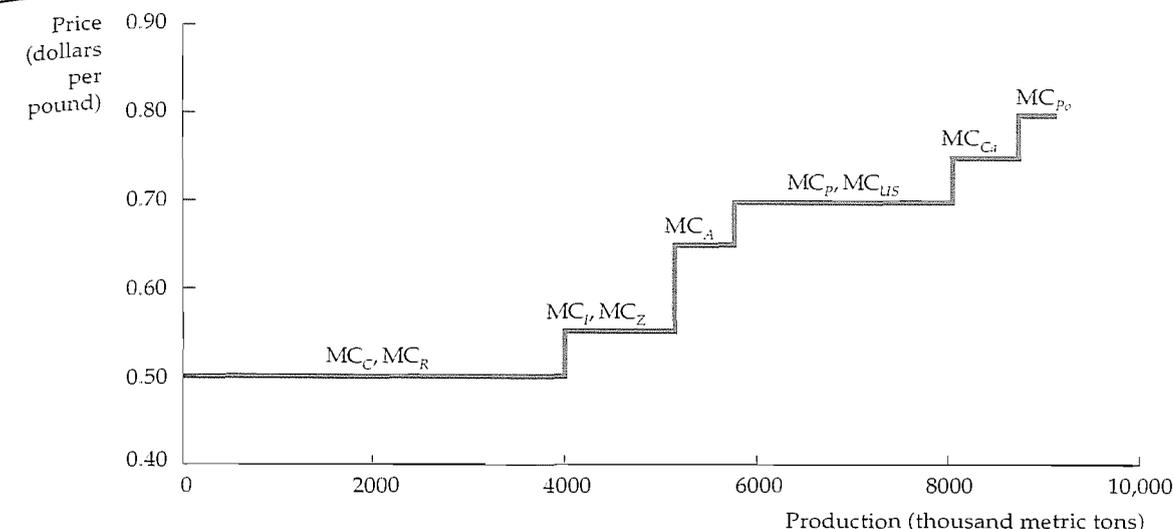
The lowest-cost copper is mined in Chile and Russia, where the marginal cost of refined copper is about 50 cents per pound.⁵ The line segment labeled MC_C, MC_R represents the marginal cost curve for these countries. The curve is horizontal until the total capacity to mine and refine copper for these two countries is reached. (That point is reached at a production level of about 4 million metric tons per year.) The line segment MC_I, MC_Z describes the marginal cost curve for Indonesia and Zambia (where the marginal cost is about 55 cents per pound). Likewise, line segment MC_A represents the marginal cost curve for Australia, and so on.

TABLE 8.1 The World Copper Industry (1999)

COUNTRY	ANNUAL PRODUCTION (THOUSAND METRIC TONS)	MARGINAL COST (DOLLARS PER POUND)
Australia	600	0.65
Canada	710	0.75
Chile	3,660	0.50
Indonesia	750	0.55
Peru	450	0.70
Poland	420	0.80
Russia	450	0.50
United States	1,850	0.70
Zambia	280	0.55

⁴ Our thanks to James Burrows, Michael Loreth, and George Rainville of Charles River Associates, Inc., who were kind enough to provide the data. The original source of the data is U.S. Geological Survey, Mineral Commodity Summaries, January 1999. Updated data and related information are available on the Web at <http://minerals.usgs.gov/minerals/pubs/commodity/copper/240399.pdf>.

⁵ We are presuming that marginal and average costs of production are approximately the same.

**FIGURE 8.10 The Short-Run World Supply of Copper**

The supply curve for world copper is obtained by summing the marginal cost curves for each of the major copper-producing countries. The supply curve slopes upward because the marginal cost of production ranges from a low of 50 cents in Chile and Russia to a high of 80 cents in Poland.

The world supply curve is obtained by summing each nation's supply curve horizontally. The slope and the elasticity of the supply curve depend on the price of copper. At relatively low prices, such as 50–55 cents per pound, the curve is quite elastic because small price increases lead to substantial increases in refined copper. But at higher prices—say, above 75 cents per pound—the supply curve becomes quite inelastic because at such prices all producers would be operating at capacity.

Producer Surplus in the Short Run

In Chapter 4, we measured consumer surplus as the difference between the maximum that a person would pay for an item and its market price. An analogous concept applies to firms. If marginal cost is rising, the price of the product is greater than marginal cost for every unit produced except the last one. As a result, firms earn a surplus on all but the last unit of output. The **producer surplus** of a firm is the sum over all units produced of the differences between the market price of the good and the marginal costs of production. Just as consumer surplus measures the area below an individual's demand curve and above the market price of the product, producer surplus measures the area above a producer's supply curve and below the market price.

Figure 8.11 illustrates short-run producer surplus for a firm. The profit-maximizing output is q^* , where $P = MC$. The surplus that the producer obtains from selling each unit is the difference between the price and the marginal cost of producing the unit. The producer surplus is then the sum of these "unit surpluses" over all units that the firm produces. It is given by the yellow area under the firm's horizontal demand curve and above its marginal cost curve, from zero output to the profit-maximizing output q^* .

For a review of consumer surplus, see §4.4, where it is defined as the difference between what a consumer is willing to pay for a good and what the consumer actually pays when buying it.

producer surplus Sum over all units produced by a firm of differences between market price of a good and marginal costs of production.

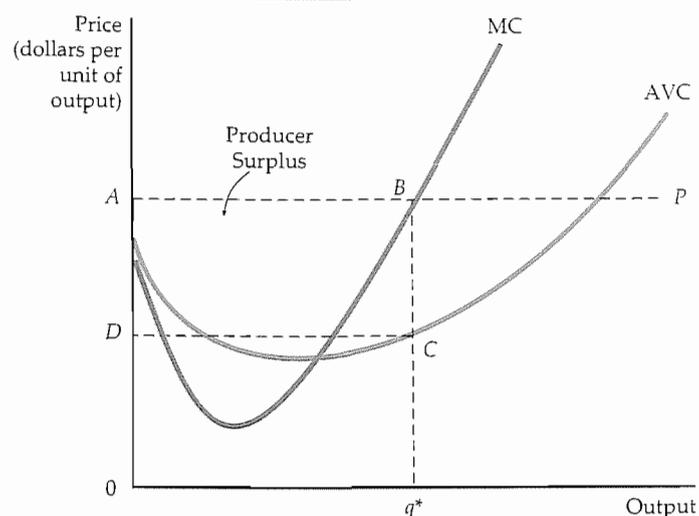


FIGURE 8.11 Producer Surplus for a Firm

The producer surplus for a firm is measured by the yellow area below the market price and above the marginal cost curve, between outputs 0 and q^* , the profit-maximizing output. Alternatively, it is equal to rectangle $ABCD$ because the sum of all marginal costs up to q^* is equal to the variable costs of producing q^* .

When we add the marginal costs of producing each level of output from 0 to q^* , we find that the sum is the total variable cost of producing q^* . Marginal cost reflects increments to cost associated with increases in output; because fixed cost does not vary with output, the sum of all marginal costs must equal the sum of the firm's variable costs.⁶ Thus producer surplus can alternatively be defined as *the difference between the firm's revenue and its total variable cost*. In Figure 8.11, producer surplus is also given by the rectangle $ABCD$, which equals revenue ($0ABq^*$) minus variable cost ($0DCq^*$).

Producer Surplus versus Profit Producer surplus is closely related to profit but is not equal to it. In the short run, producer surplus is equal to revenue minus variable cost, which is *variable profit*. Total profit, on the other hand, is equal to revenue minus *all* costs, both variable and fixed:

$$\begin{aligned} \text{Producer surplus} = \text{PS} &= R - \text{VC} \\ \text{Profit} = \pi &= R - \text{VC} - \text{FC} \end{aligned}$$

It follows that in the short run, when fixed cost is positive, producer surplus is greater than profit.

The extent to which firms enjoy producer surplus depends on their costs of production. Higher-cost firms have less producer surplus, and lower-cost firms have more. By adding up the producer surpluses of all firms, we can determine the producer surplus for a market. This can be seen in Figure 8.12. The market

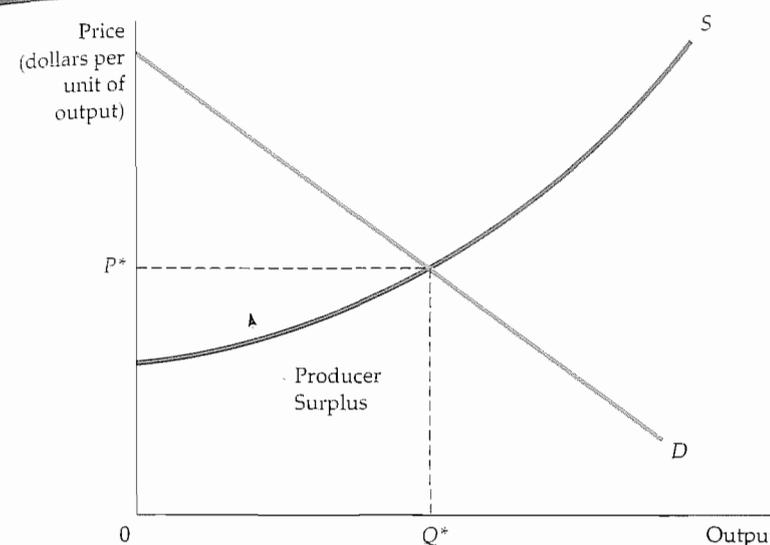


FIGURE 8.12 Producer Surplus for a Market

The producer surplus for a market is the area below the market price and above the market supply curve, between 0 and output Q^* .

supply curve begins at the vertical axis at a point representing the average variable cost of the lowest-cost firm in the market. Producer surplus is the area that lies below the market price of the product and above the supply curve between the output levels 0 and Q^* .

8.7 Choosing Output in the Long Run

In the long run, a firm can alter all its inputs, including plant size. It can decide to shut down (i.e., to *exit* the industry) or to begin producing a product for the first time (i.e., to *enter* an industry). Because we are concerned here with competitive markets, we allow for *free entry* and *free exit*. In other words, we are assuming that firms may enter or exit without any legal restriction or any special costs associated with entry. (Recall from Section 8.1 that this is one of the key assumptions underlying perfect competition.) After analyzing the long-run output decision of a profit-maximizing firm in a competitive market, we discuss the nature of competitive equilibrium in the long run. We also discuss the relationship between entry and economic and accounting profits.

Long-Run Profit Maximization

Figure 8.13 shows how a competitive firm makes its long-run, profit-maximizing output decision. As in the short run, it faces a horizontal demand curve. (In Figure 8.13 the firm takes the market price of \$40 as given.) Its short-run average (total) cost curve SAC and short-run marginal cost curve SMC are low enough for the firm to make a positive profit, given by rectangle $ABCD$, by producing an output of q_1 , where $SMC = P = MR$. The long-run average cost curve LAC

⁶ The area under the marginal cost curve from 0 to q^* is $TC(q^*) - TC(0) = TC - FC = VC$.

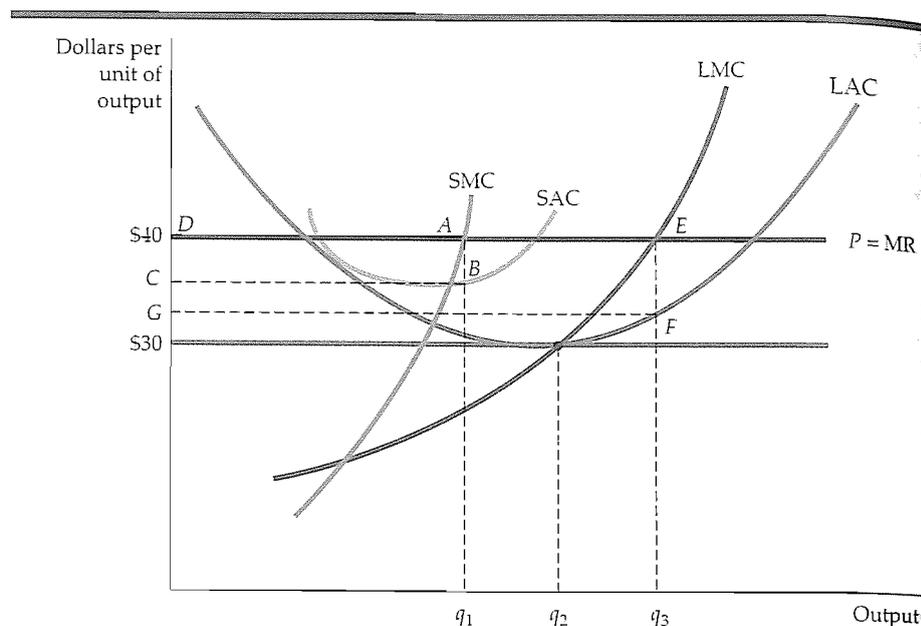


FIGURE 8.13 Output Choice in the Long Run

The firm maximizes its profit by choosing the output at which price equals long-run marginal cost LMC. In the diagram, the firm increases its profit from $ABCD$ to $EFGD$ by increasing its output in the long run.

In §7.4, we explain that economies of scale arise when a firm can double its output for less than twice the cost.

reflects the presence of economies of scale up to output level q_2 and diseconomies of scale at higher output levels. The long-run marginal cost curve LMC cuts the long-run average cost from below at q_2 , the point of minimum long-run average cost.

If the firm believes the market price will remain at \$40, it will want to increase the size of its plant to produce at output q_3 at which its long-run marginal cost equals the \$40 price. When this expansion is complete, the profit margin will increase from AB to EF , and total profit will increase from $ABCD$ to $EFGD$. Output q_3 is profit-maximizing for the firm because at any lower output (say q_2), the marginal revenue from additional production is greater than the marginal cost. Expansion is, therefore, desirable. But at any output greater than q_3 , marginal cost is greater than marginal revenue. Additional production would therefore reduce profit. In summary, *the long-run output of a profit-maximizing competitive firm is the point at which long-run marginal cost equals the price.*

Note that the higher the market price, the higher the profit that the firm can earn. Correspondingly, as the price of the product falls from \$40 to \$30, the profit also falls. At a price of \$30, the firm's profit-maximizing output is q_2 , the point of long-run minimum average cost. In this case, because $P = ATC$, the firm earns zero economic profit.

Long-Run Competitive Equilibrium

For an equilibrium to arise in the long run, certain economic conditions must prevail. Firms in the market must have no desire to withdraw at the same time that no firms outside the market wish to enter. But what is the exact relationship

between profitability, entry, and long-run competitive equilibrium? The answer can be seen by relating economic profit to the incentive to enter and exit a market.

Accounting Profit and Economic Profit As we saw in Chapter 7, it is important to distinguish between accounting profit and economic profit. Accounting profit is measured by the difference between the firm's revenues and its cash flows for labor, raw materials, and interest plus depreciation expenses. Economic profit takes into account opportunity costs. One such opportunity cost is the return to the firm's owners if their capital were used elsewhere. Suppose, for example, that the firm uses labor and capital inputs; its capital equipment has been purchased. Accounting profit will equal revenues R minus labor costs wL , which is positive. However, economic profit π equals revenues R minus labor cost wL minus the capital cost, rK :

$$\pi = R - wL - rK$$

As we explained in Chapter 7, the correct measure of capital cost is the user cost of capital, which is the annual return the firm could earn by investing its money elsewhere instead of purchasing capital, plus the annual depreciation on the capital.

Zero Economic Profit When a firm goes into a business, it does so in the expectation that it will earn a return on its investment. A **zero economic profit** means that the firm is earning a *normal*—i.e., competitive—return on that investment. This normal return, which is part of the user cost of capital, is the firm's opportunity cost of using its money to buy capital rather than investing it elsewhere. Thus, *a firm earning zero economic profit is doing as well by investing its money in capital as it could by investing elsewhere*—it is earning a competitive return on its money. Such a firm, therefore, is performing adequately, and should stay in business. (A firm earning a *negative* economic profit, however, should consider going out of business if it does not expect to improve its financial picture.)

As we will see, in competitive markets economic profit becomes zero in the long run. Zero economic profit signifies not that firms are performing poorly, but rather that the industry is competitive.

Entry and Exit Figure 8.13 shows how a \$40 price induces a firm to increase output and realize a positive profit. Because profit is calculated after subtracting the opportunity cost of capital, a positive profit means an unusually high return on a financial investment, which can be earned by entering a profitable industry. This high return causes investors to direct resources away from other industries and into this one—there will be *entry* into the market. Eventually the increased production associated with new entry causes the market supply curve to shift to the right. As a result, market output increases and the market price of the product falls.⁷ Figure 8.14 illustrates this. In part (b) of the figure, the supply curve has shifted from S_1 to S_2 , causing the price to fall from P_1 (\$40) to P_2 (\$30). In part (a), which applies to a single firm, the long-run average cost curve is tangent to the horizontal price line at output q_2 .

⁷ We discuss why the long-run supply curve might be upward sloping in the next section.

zero economic profit A firm is earning a normal return on its investment—i.e., it is doing as well as it could by investing its money elsewhere.

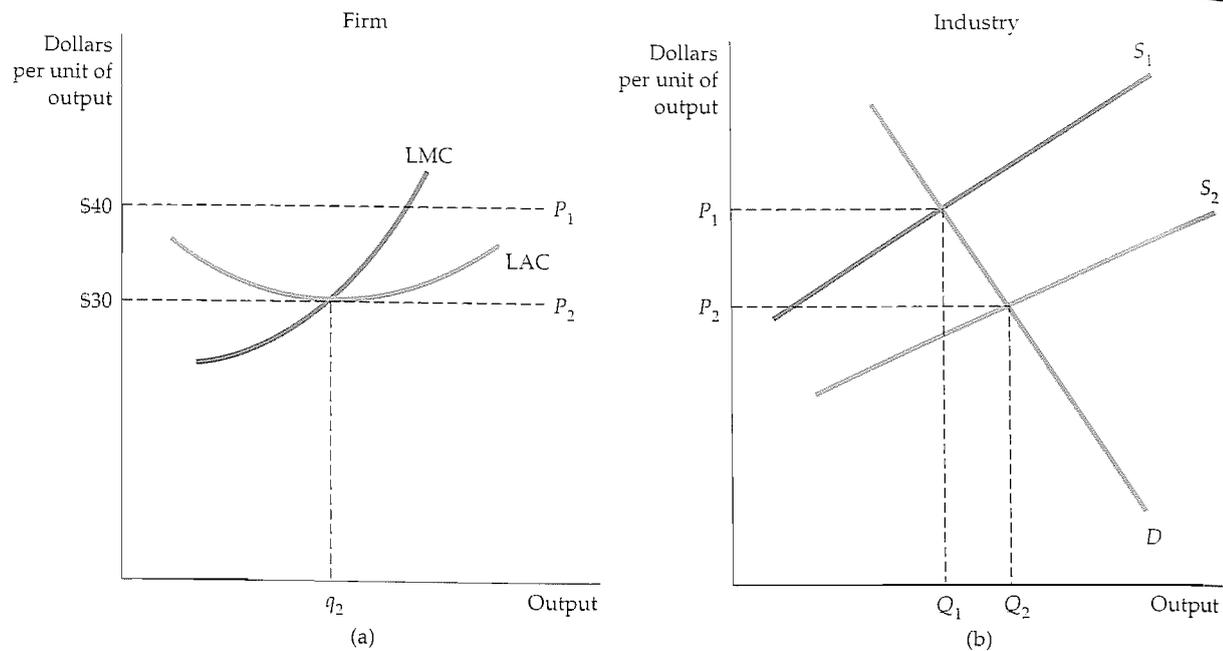


FIGURE 8.14 Long-Run Competitive Equilibrium

Initially the long-run equilibrium price of a product is \$40 per unit, as shown in (b) as the intersection of demand curve D and supply curve S_1 . In (a) we see that firms earn positive profits because long-run average cost reaches a minimum of \$30 (at q_2). This positive profit encourages entry of new firms and causes a shift to the right in the supply curve to S_2 as shown in (a). The long-run equilibrium occurs at a price of \$30 as shown in (b), where each firm earns zero profit and there is no incentive to enter or exit the industry.

long-run competitive equilibrium All firms in an industry are maximizing profit, no firm has an incentive to enter or exit, and price is such that quantity supplied equals quantity demanded.

When a firm earns zero economic profit, it has no incentive to exit the industry. Likewise, other firms have no special incentive to enter. A **long-run competitive equilibrium** occurs when three conditions hold:

1. All firms in the industry are maximizing profit.
2. No firm has an incentive either to enter or exit the industry because all firms are earning zero economic profit.
3. The price of the product is such that the quantity supplied by the industry is equal to the quantity demanded by consumers.

The dynamic process that leads to long-run equilibrium creates a puzzle. Firms enter the market because they hope to earn a profit, and likewise they exit because of economic losses. In long-run equilibrium, however, firms earn zero economic profit. Why does a firm enter a market knowing that it will eventually earn zero profit? The answer is that zero economic profit represents a competitive return for the firm's investment of financial capital. With zero economic profit, the firm has no incentive to go elsewhere because it cannot do better financially by doing so. If the firm happens to enter a market sufficiently early to enjoy an economic profit in the short run, so much the better. Similarly, if a firm exits an unprofitable market quickly, it can save its investors money. Thus the concept of long-run equilibrium tells us the direction that firms' behavior is likely to take. The idea of an eventual zero-profit, long-run equilibrium should not discourage a manager—it should be seen in a positive light, because it reflects the opportunity to earn a competitive return.

Firms Having Identical Costs To see why all the conditions for long-run equilibrium must hold, assume that all firms have identical costs. Now consider what happens if too many firms enter the industry in response to an opportunity for profit. The industry supply curve in Figure 8.14(b) will shift further to the right, and price will fall below \$30—say, to \$25. At that price, however, firms will lose money. As a result, some firms will exit the industry. Firms will continue to exit until the market supply curve shifts back to S_2 . Only when there is no incentive to exit or enter can a market be in long-run equilibrium.

Firms Having Different Costs Now suppose that all firms in the industry do not have identical cost curves. Perhaps one firm has a patent that lets it produce at a lower average cost than all other firms. In that case, it is consistent with long-run equilibrium for that firm to earn a greater *accounting* profit and to enjoy a higher producer surplus than other firms. As long as other investors and firms cannot acquire the patent that lowers costs, they have no incentive to enter the industry. Conversely, as long as the process is particular to this product and this industry, the fortunate firm has no incentive to exit the industry.

The distinction between accounting profit and economic profit is important here. If the patent is profitable, other firms in the industry will pay to use it (or attempt to buy the entire firm to acquire it). The increased value of the patent thus represents an opportunity cost to the firm that holds it. It could sell the rights to the patent rather than use it. If all firms are equally efficient otherwise, the *economic* profit of the firm falls to zero. However, if the firm with the patent is more efficient than other firms, then it will be earning a positive profit. But if the patent holder is otherwise less efficient, it should sell off the patent and exit the industry.

The Opportunity Cost of Land There are other instances in which firms earning positive accounting profit may be earning zero economic profit. Suppose, for example, that a clothing store happens to be located near a large shopping center. The additional flow of customers may substantially increase the store's accounting profit because the cost of the land is based on its historical cost. However, as far as economic profit is concerned, the cost of the land should reflect its opportunity cost, which in this case is the current market value of the land. When the opportunity cost of land is included, the profitability of the clothing store is no higher than that of its competitors.

Thus the condition that economic profit be zero is essential for the market to be in long-run equilibrium. By definition, positive economic profit represents an opportunity for investors and an incentive to enter an industry. Positive accounting profit, however, may signal that firms already in the industry possess valuable assets, skills, or ideas, which will not necessarily encourage entry.

Economic Rent

We have seen that some firms earn higher accounting profit than others because they have access to factors of production that are in limited supply; these might include land and natural resources, entrepreneurial skill, or other creative talent. In these situations what makes economic profit zero in the long run is the willingness of other firms to use the factors of production that are in limited supply. The positive accounting profits are therefore translated into *economic rent* that is earned by the scarce factors. **Economic rent** is what firms are willing to pay for an input less the minimum amount necessary to buy it. In competitive markets, in both the short and the long run, economic rent is often positive even though profit is zero.

economic rent Amount that firms are willing to pay for an input less the minimum amount necessary to obtain it.

For example, suppose that two firms in an industry own their land outright, thus the minimum cost of obtaining the land is zero. One firm, however, is located on a river and can ship its products for \$10,000 a year less than the other firm, which is inland. In this case, the \$10,000 higher profit of the first firm is due to the \$10,000 per year economic rent associated with its river location. The rent is created because the land along the river is valuable and other firms would be willing to pay for it. Eventually, the competition for this specialized factor of production will increase the value of that factor to \$10,000. Land rent—the difference between \$10,000 and the zero cost of obtaining the land—is also \$10,000. Note that while the economic rent has increased, the economic profit of the firm on the river has become zero.

The presence of economic rent explains why there are some markets in which firms cannot enter in response to profit opportunities. In those markets, the supply of one or more inputs is fixed, one or more firms earn economic rents, and all firms enjoy zero economic profit. Zero economic profit tells a firm that it should remain in a market only if it is at least as efficient in production as other firms. It also tells possible entrants to the market that entry will be profitable only if they can produce more efficiently than firms already in the market.

Producer Surplus in the Long Run

Suppose that a firm is earning a positive accounting profit but that there is no incentive for other firms to enter or exit the industry. This profit must reflect economic rent. How then does rent relate to producer surplus? To begin with, note that while economic rent applies to factor inputs, producer surplus applies to outputs. Note also that producer surplus measures the difference between the market price a producer receives and the marginal cost of production. Thus, in the long run, in a competitive market, *the producer surplus that a firm earns on the output that it sells consists of the economic rent that it enjoys from all its scarce inputs.*⁸

Let's say, for example, that a baseball team has a franchise allowing it to operate in a particular city. Suppose also that the only alternative location for the team is a city in which the team will generate substantially lower revenues. The team will therefore earn an economic rent associated with its current location. This rent will reflect the difference between what the firm would be willing to pay for its current location and the amount needed to locate in the alternative city. The firm will also be earning a producer surplus associated with the sale of baseball tickets and other franchise items at its current location. This surplus will reflect all economic rents, including those rents associated with the firm's other factor inputs (the stadium and the players).

Figure 8.15 shows that firms earning economic rent earn the same economic profit as firms that do not earn rent. Part (a) shows the economic profit of a baseball team located in a moderate-sized city. The average price of a ticket is \$7, and costs are such that the team earns zero economic profit. Part (b) shows the profit of a team with the same costs, even though it is located in a larger city. Because more people want to see baseball games, the latter team can sell tickets for \$10 apiece and thereby earn an accounting profit of \$3 on each ticket. However, the rent associated with the more desirable location represents a cost to the firm—an opportunity cost—because it could sell its franchise to another team. As a result, the economic profit in the larger city is also zero.

⁸ In a noncompetitive market, producer surplus will reflect economic profit as well as economic rent.

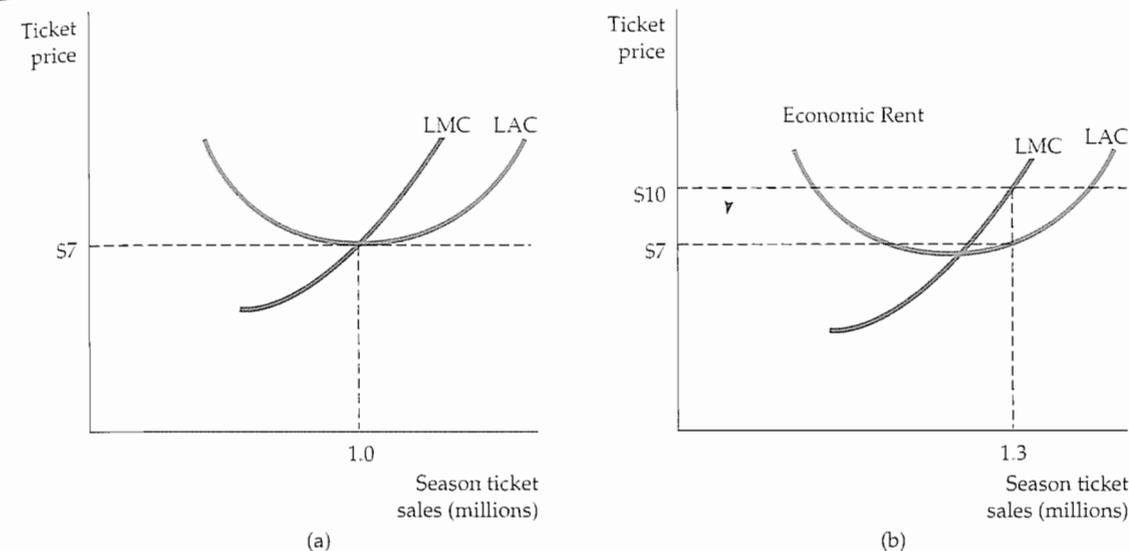


FIGURE 8.15 Firms Earn Zero Profit in Long-Run Equilibrium

In long-run equilibrium, all firms earn zero economic profit. In (a), a baseball team in a moderate-sized city sells enough tickets so that price (\$7) is equal to marginal and average cost. In (b), the demand is greater, so a \$10 price can be charged. The team increases sales to the point at which the average cost of production plus the average economic rent is equal to the ticket price. When the opportunity cost associated with owning the franchise is taken into account, the team earns zero economic profit.

8.8 The Industry's Long-Run Supply Curve

In our analysis of short-run supply, we first derived the firm's supply curve and then showed how the summation of individual firms' supply curves generated a market supply curve. We cannot, however, analyze long-run supply in the same way: In the long run, firms enter and exit markets as the market price changes. This makes it impossible to sum up supply curves—we do not know which firms' supplies to add up to get market totals.

The shape of the long-run supply curve depends on the extent to which increases and decreases in industry output affect the prices that firms must pay for inputs into the production process. To determine long-run supply, we assume all firms have access to the available production technology. Output is increased by using more inputs, not by invention. We also assume that the conditions underlying the market for inputs to production do not change when the industry expands or contracts. For example, an increased demand for labor does not increase a union's ability to negotiate a better wage contract for its workers.

In our analysis of long-run supply, it will be useful to distinguish among three types of industries: *constant-cost*, *increasing-cost*, and *decreasing-cost*.

Constant-Cost Industry

Figure 8.16 shows the derivation of the long-run supply curve for a **constant-cost industry**. A firm's output choice is given in (a), while industry output is shown in (b). Assume that the industry is initially in equilibrium at the intersection of

constant-cost industry
Industry whose long-run supply curve is horizontal.

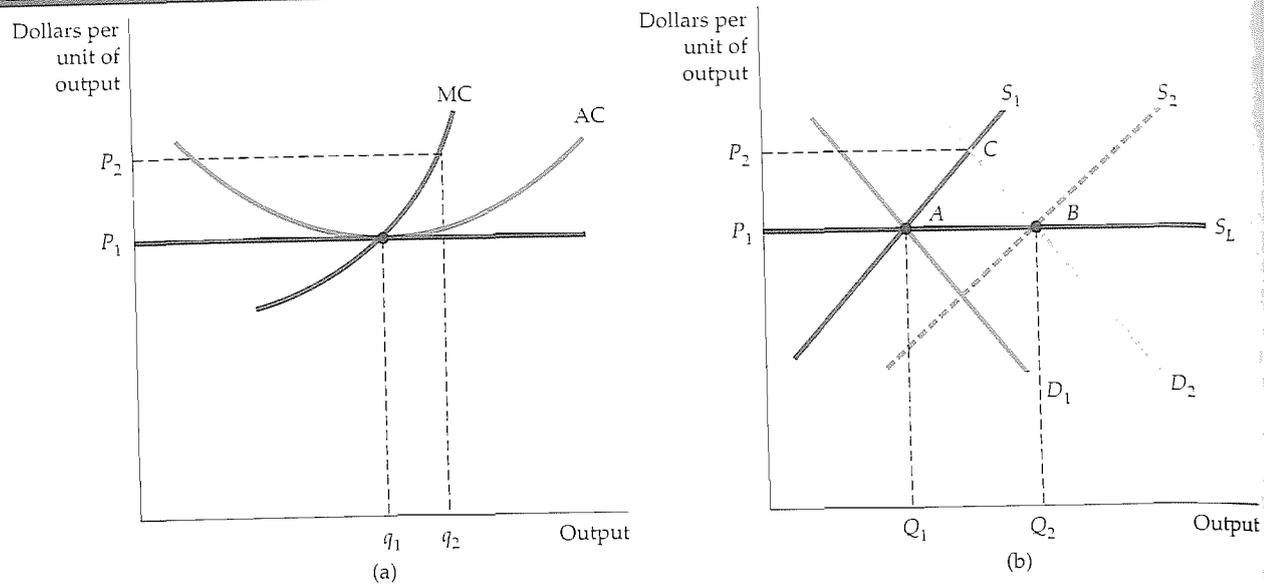


FIGURE 8.16 Long-Run Supply in a Constant-Cost Industry

In (b), the long-run supply curve in a constant-cost industry is a horizontal line S_L . When demand increases, initially causing a price rise (represented by a move from point A to point C), the firm initially increases its output from q_1 to q_2 , as shown in (a). But the entry of new firms causes a shift to the right in industry supply. Because input prices are unaffected by the increased output of the industry, entry occurs until the original price is obtained (at point B).

market demand curve D_1 and short-run market supply curve S_1 . Point A at the intersection of demand and supply is on the long-run supply curve S_L because it tells us that the industry will produce Q_1 units of output when the long-run equilibrium price is P_1 .

To obtain other points on the long-run supply curve, suppose the market demand for the product unexpectedly increases (say, because of a reduction in personal income taxes). A typical firm is initially producing at an output of q_1 , where P_1 is equal to long-run marginal cost and long-run average cost. But because the firm is also in short-run equilibrium, price also equals short-run marginal cost. Suppose that the tax cut shifts the market demand curve from D_1 to D_2 . Demand curve D_2 intersects supply curve S_1 at C. As a result, price increases from P_1 to P_2 .

Part (a) of Figure 8.16 shows how this price increase affects a typical firm in the industry. When the price increases to P_2 , the firm follows its short-run marginal cost curve and increases output to q_2 . This output choice maximizes profit because it satisfies the condition that price equal short-run marginal cost. If every firm responds this way, each will be earning a positive profit in short-run equilibrium. This profit will be attractive to investors and will cause existing firms to expand operations and new firms to enter the market.

As a result, in Figure 8.16(b) the short-run supply curve shifts to the right from S_1 to S_2 . This shift causes the market to move to a new long-run equilibrium at the intersection of D_2 and S_2 . For this intersection to be a long-run equilibrium, output must expand enough so that firms are earning zero profit and the incentive to enter or exit the industry disappears.

In a constant-cost industry, the additional inputs necessary to produce higher output can be purchased without an increase in per-unit price. This might happen, for example, if unskilled labor is a major input in production, and the market wage of unskilled labor is unaffected by the increase in the demand for labor.

Because the prices of inputs have not changed, firms' cost curves are also unchanged; the new equilibrium must be at a point such as B in Figure 8.16(b), at which price is equal to P_1 , the original price before the unexpected increase in demand occurred.

The long-run supply curve for a constant-cost industry is, therefore, a horizontal line at a price that is equal to the long-run minimum average cost of production. At any higher price, there would be positive profit, increased entry, increased short-run supply, and thus downward pressure on price. Remember that in a constant-cost industry, input prices do not change when conditions change in the output market. Constant-cost industries can have horizontal long-run average cost curves.

Increasing-Cost Industry

In an increasing-cost industry, the prices of some or all inputs to production increase as the industry expands and the demand for the inputs grows. This situation might arise, for example, if the industry uses skilled labor, which becomes in short supply as the demand for it increases. If a firm requires mineral resources that are available only on certain types of land, the cost of land as an input increases with output. Figure 8.17 shows the derivation of long-run supply, which is similar to the previous constant-cost derivation. The industry is initially in equilibrium at A in part (b). When the demand curve unexpectedly shifts from D_1 to D_2 , the price of the product increases in the short run to P_2 , and industry output increases from Q_1 to Q_2 . A typical firm, as shown in part (a), increases its output from q_1 to q_2 in response to the higher price by moving along its short-run marginal cost curve. The higher profit earned by this and other firms induces new firms to enter the industry.

increasing-cost industry
Industry whose long-run supply curve is upward sloping.

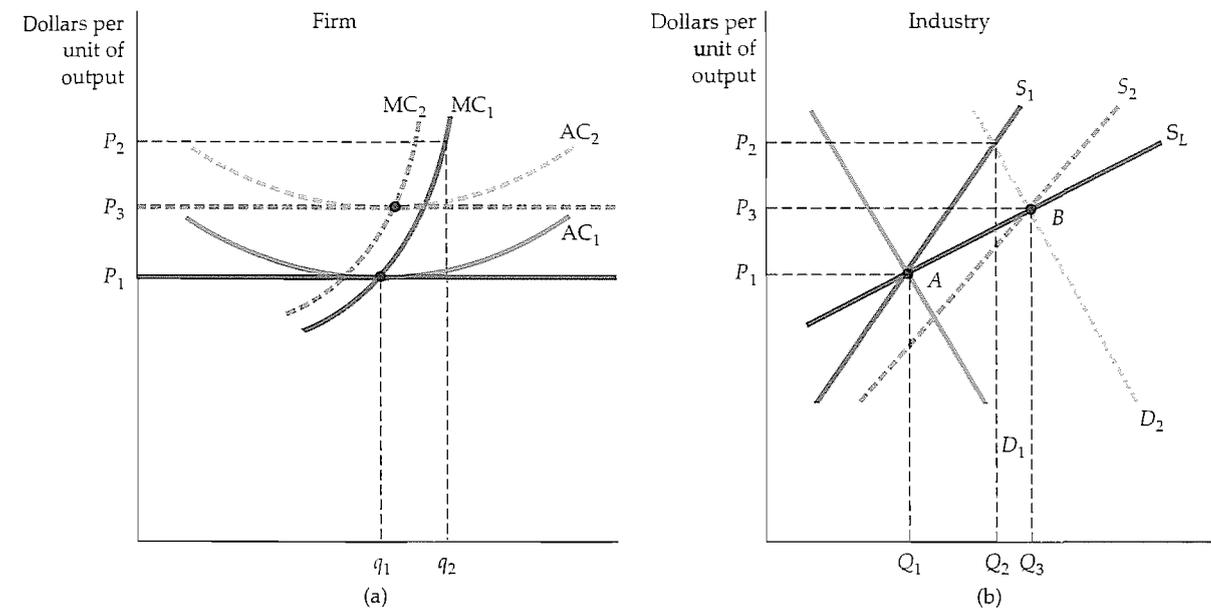


FIGURE 8.17 Long-Run Supply in an Increasing-Cost Industry

In (b), the long-run supply curve in an increasing-cost industry is an upward-sloping curve S_L . When demand increases, initially causing a price rise, the firms increase their output from q_1 to q_2 in (a). In that case, the entry of new firms causes a shift to the right in supply. Because input prices increase as a result, the new long-run equilibrium occurs at a higher price than the initial equilibrium.

As new firms enter and output expands, increased demand for inputs causes some or all input prices to increase. The short-run market supply curve shifts to the right as before, though not as much, and the new equilibrium at B results in a price P_3 that is higher than the initial price P_1 . Because the higher input prices raise the firms' short-run and long-run cost curves, the higher market price is needed to ensure that firms earn zero profit in long-run equilibrium. Figure 8.17(a) illustrates this. The average cost curve shifts up from AC_1 to AC_2 , while the marginal cost curve shifts to the left from MC_1 to MC_2 . The new long-run equilibrium price P_3 is equal to the new minimum average cost. As in the constant-cost case, the higher short-run profit caused by the initial increase in demand disappears in the long run as firms increase output and input costs rise.

The new equilibrium at B in Figure 8.17(b) is, therefore, on the long-run supply curve for the industry. In an increasing-cost industry, the long-run industry supply curve is upward sloping. The industry produces more output, but only at the higher price needed to compensate for the increase in input costs. The term "increasing cost" refers to the upward shift in the firms' long-run average cost curves, not to the positive slope of the cost curve itself.

Decreasing-Cost Industry

The industry supply curve can also be downward sloping. In this case, the unexpected increase in demand causes industry output to expand as before. But as the industry grows larger, it can take advantage of its size to obtain some of its inputs more cheaply. For example, a larger industry may allow for an improved transportation system or for a better, less expensive financial network. In this case, firms' average cost curves shift downward (even though they do not enjoy economies of scale), and the market price of the product falls. The lower market price and lower average cost of production induce a new long-run equilibrium with more firms, more output, and a lower price. Therefore, in a **decreasing-cost industry**, the long-run supply curve for the industry is downward sloping.

decreasing-cost industry
Industry whose long-run supply curve is downward sloping.

The Effects of a Tax

In Chapter 6, we saw that a tax on one of a firm's inputs (in the form of an effluent fee) creates an incentive for the firm to change the way it uses inputs in its production process. Now we consider ways in which a firm responds to a tax on its output. To simplify the analysis, assume that the firm uses a fixed-proportions production technology. If the firm is a polluter, the output tax might encourage the firm to reduce its output, and therefore its effluent, or it might be imposed merely to raise revenue.

First, suppose the output tax is imposed only on this firm and thus does not affect the market price of the product. We will see that the tax on output encourages the firm to reduce its output. Figure 8.18 shows the relevant short-run cost curves for a firm enjoying positive economic profit by producing an output of q_1 and selling its product at the market price P_1 . Because the tax is assessed for every unit of output, it raises the firm's marginal cost curve from MC_1 to $MC_2 = MC_1 + t$, where t is the tax per unit of the firm's output. The tax also raises the average variable cost curve by the amount t .

The output tax can have two possible effects. If the firm can still earn a positive or zero economic profit after the imposition of the tax, it will maximize its profit by choosing an output level at which marginal cost plus the tax is equal to the price of the product. Its output falls from q_1 to q_2 , and the implicit effect of the tax is to shift its supply curve upward (by the amount of the tax). If the firm can no longer earn an economic profit after the tax has been imposed, the firm will choose to exit the market.

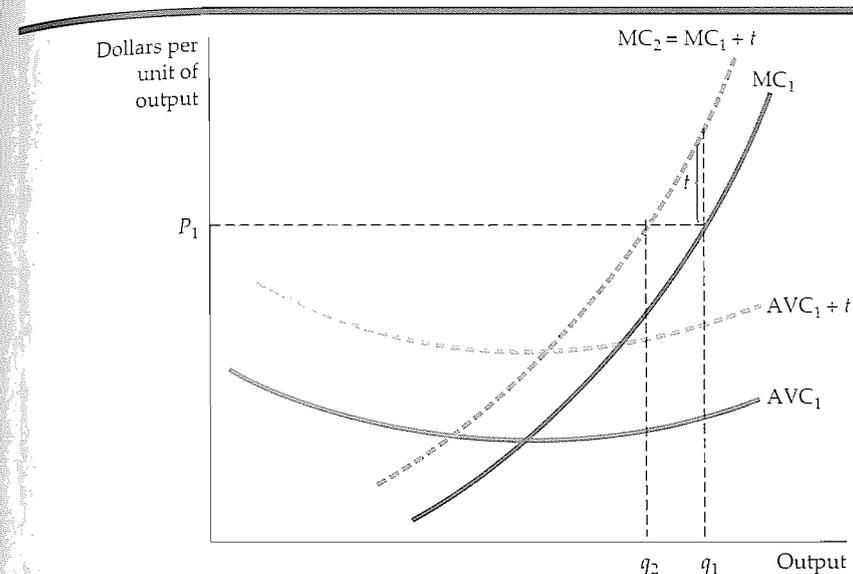


FIGURE 8.18 Effect of an Output Tax on a Competitive Firm's Output

An output tax raises the firm's marginal cost curve by the amount of the tax. The firm will reduce its output to the point at which the marginal cost plus the tax is equal to the price of the product.

Now suppose that all firms in the industry are taxed and so have increasing marginal costs. Because each firm reduces its output at the current market price, the total output supplied by the industry will also fall, causing the price of the product to increase. Figure 8.19 illustrates this. An upward shift in the supply curve, from S_1 to $S_2 = S_1 + t$, causes the market price of the product to increase (by less than the amount of the tax) from P_1 to P_2 . This increase in price diminishes some of the effects that we described previously. Firms will reduce their output less than they would without a price increase.

Finally, output taxes may also encourage some firms (those whose costs are somewhat higher than others) to exit the industry. In the process, the tax raises the long-run average cost curve for each firm.

Long-Run Elasticity of Supply

The long-run elasticity of industry supply is defined in the same way as the short-run elasticity: It is the percentage change in output ($\Delta Q/Q$) that results from a percentage change in price ($\Delta P/P$). In a constant-cost industry, the long-run supply curve is horizontal, and the long-run supply elasticity is infinitely large. (A small increase in price will induce an extremely large increase in output.) In an increasing-cost industry, however, the long-run supply elasticity will be positive but finite. Because industries can adjust and expand in the long run, we would generally expect long-run elasticities of supply to be larger than short-run elasticities.⁹ The magnitude of the elasticity will depend on the extent to

⁹ In some cases the opposite is true. Consider the elasticity of supply of scrap metal from a durable good like copper. Recall from Chapter 2 that because there is an existing stock of scrap, the long-run elasticity of supply will be *smaller* than the short-run elasticity.

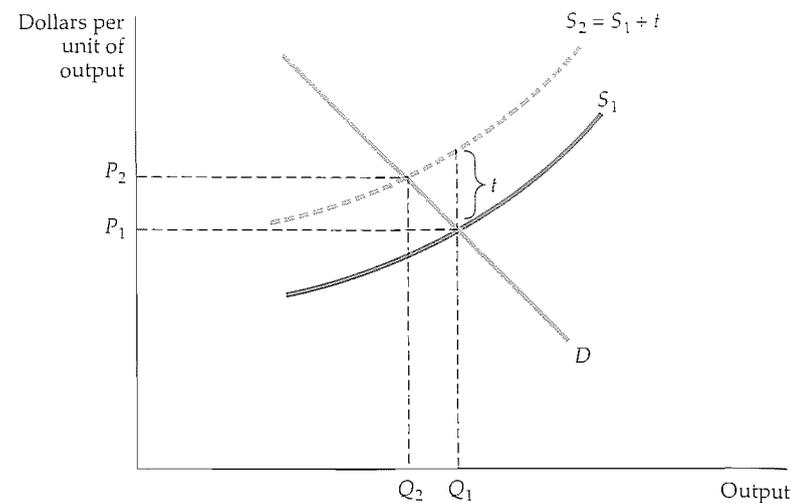


FIGURE 8.19 Effect of an Output Tax on Industry Output

An output tax placed on all firms in a competitive market shifts the supply curve for the industry upward by the amount of the tax. This shift raises the market price of the product and lowers the total output of the industry.

which input costs increase as the market expands. For example, an industry that depends on inputs that are widely available will have a more elastic long-run supply than will an industry that uses inputs in short supply.

EXAMPLE 8.5 The Long-Run Supply of Housing

Owner-occupied and rental housing provide interesting examples of the range of possible supply elasticities. People buy or rent housing to obtain the services that a house provides—a place to eat and sleep, comfort, and so on. If the price of housing services were to rise in one area of the country, the quantity of services provided could increase substantially.

To begin, consider the supply of owner-occupied housing in suburban or rural areas where land is not scarce. In this case, the price of land does not increase substantially as the quantity of housing supplied increases. Likewise, costs associated with construction are not likely to increase because there is a national market for lumber and other materials. Therefore, the long-run elasticity of the housing supply is likely to be very large, approximating a constant-cost industry. In fact, many studies find the long-run supply curve to be nearly horizontal.¹⁰

Even when elasticity of supply is measured within urban areas, where land costs rise as the demand for housing services increases, the long-run elasticity of supply is still likely to be large because land costs make up only about one-

quarter of total housing costs. In one study of urban housing supply, the price elasticity was found to be 5.3.¹¹

The market for rental housing is different, however. The construction of rental housing is often restricted by local zoning laws. Many communities outlaw it entirely, while others limit to it certain areas. Because urban land on which most rental housing is located is restricted and valuable, the long-run elasticity of supply of rental housing is much lower than the elasticity of supply of owner-occupied housing. As the price of rental housing services rises, new high-rise rental units are built and older units are renovated—a practice that increases the quantity of rental services. With urban land becoming more valuable as housing density increases, and with the cost of construction soaring with the height of buildings, increased demand causes the input costs of rental housing to rise. In this increasing-cost case, the elasticity of supply can be much less than 1; in one study, the authors found the supply elasticity to be between 0.3 and 0.7.¹²

SUMMARY

1. Managers can operate in accordance with a complex set of objectives and under various constraints. However, we can assume that firms act as if they are maximizing long-run profit.
2. Many markets may approximate perfect competition in that one or more firms act as if they face a nearly horizontal demand curve. In general, the number of firms in an industry is not always a good indicator of the extent to which that industry is competitive.
3. Because a firm in a competitive market has a small share of total industry output, it makes its output choice under the assumption that its production decision will have no effect on the price of the product. In this case, the demand curve and the marginal revenue curve are identical.
4. In the short run, a competitive firm maximizes its profit by choosing an output at which price is equal to (short-run) marginal cost. Price must, however, be greater than or equal to the firm's minimum average variable cost of production.
5. The short-run market supply curve is the horizontal summation of the supply curves of the firms in an industry. It can be characterized by the elasticity of supply: the percentage change in quantity supplied in response to a percentage change in price.
6. The producer surplus for a firm is the difference between its revenue and the minimum cost that would be necessary to produce the profit-maximizing output. In both the short run and the long run, producer surplus is the area under the horizontal price line and above the marginal cost of production.
7. *Economic rent* is the payment for a scarce factor of production less the minimum amount necessary to hire that factor. In the long run in a competitive market, producer surplus is equal to the economic rent generated by all scarce factors of production.
8. In the long run, profit-maximizing competitive firms choose the output at which price is equal to long-run marginal cost.
9. A long-run competitive equilibrium occurs under these conditions: (a) when firms maximize profit; (b) when all firms earn zero economic profit, so that there is no incentive to enter or exit the industry; and (c) when the quantity of the product demanded is equal to the quantity supplied.
10. The long-run supply curve for a firm is horizontal when the industry is a constant-cost industry in which the increased demand for inputs to production (associated with an increased demand for the product) has no effect on the market price of the inputs. But the long-run supply curve for a firm is upward sloping in an increasing-cost industry, where the increased demand for inputs causes the market price of some or all inputs to rise.

¹⁰ For a review of the relevant literature, see Dixie M. Blackley, "The Long-Run Elasticity of New Housing Supply in the United States: Empirical Evidence for 1950 to 1994," *Journal of Real Estate Finance and Economics* 18(1999): 25–42.

¹¹ See Barton A. Smith, "The Supply of Urban Housing," *Journal of Political Economy* 40, no. 3 (August 1976): 389–405.

¹² See Frank deLeeuw and Nkanta Ekanem, "The Supply of Rental Housing," *American Economic Review* 61 (December 1971): 806–17, table 5.2.

QUESTIONS FOR REVIEW

- Why would a firm that incurs losses choose to produce rather than shut down?
- The supply curve for a firm in the short run is the short-run marginal cost curve (above the point of minimum average variable cost). Why is the supply curve in the long run *not* the long-run marginal cost curve (above the point of minimum average total cost)?
- In long-run equilibrium, all firms in the industry earn zero economic profit. Why is this true?
- What is the difference between economic profit and producer surplus?
- Why do firms enter an industry when they know that in the long run, economic profit will be zero?
- At the beginning of the twentieth century, there were many small American automobile manufacturers. At the end of the century, there are only two large ones. Suppose that this situation is not the result of lax federal enforcement of antimonopoly laws. How do you explain the decrease in the number of manufacturers? (*Hint*: What is the inherent cost structure of the automobile industry?)
- Because Industry X is characterized by perfect competition, every firm is earning zero economic profit. If the product price falls, no firms can survive. Do you agree or disagree? Discuss.
- An increase in the demand for video films also increases the salaries of actors and actresses. Is the long-run supply curve for films likely to be horizontal or upward sloping? Explain.
- True or false: A firm should always produce at an output at which long-run average cost is minimized. Explain.
- Can there be constant returns to scale in an industry characterized by an upward-sloping supply curve? Explain.
- What assumptions are necessary for a market to be perfectly competitive? In light of what you have learned in this chapter, why is each of these assumptions important?
- Suppose a competitive industry faces an increase in demand (i.e., the curve shifts upward). What are the steps by which a competitive market ensures increased output? Does your answer change if the government imposes a price ceiling?
- The government passes a law that allows a substantial subsidy for every acre of land used to grow tobacco. How does this program affect the long-run supply curve for tobacco?

EXERCISES

- Using data in Table 8.A on the following page, explain what happens to a firm's output choice and profit if the price of the product falls from \$40 to \$35.
- Again, using the data in the table, show what happens to the firm's output choice and profit if the fixed cost of production increases from \$50 to \$100 and then to \$150. What general conclusion can you reach about the effects of fixed costs on output choice?
- Suppose you are the manager of a watchmaking firm operating in a competitive market. Your cost of production is given by $C = 100 + Q^2$, where Q is the level of output and C is total cost. The marginal cost of production is $2Q$. The fixed cost of production is \$100.
 - If the price of watches is \$60, how many watches should you produce to maximize profit?
 - What will your profit level be?
 - At what minimum price will you produce a positive output?
- Use the information in Table 8.A to answer the following.
 - Derive the firm's short-run supply curve. (*Hint*: You may want to plot the appropriate cost curves.)
 - If 100 identical firms are in the market, what is the industry supply curve?
- A sales tax of \$1 per unit of output is placed on one firm whose product sells for \$5 in a competitive industry.
 - How will this tax affect the cost curves for the firm?
 - What will happen to price, output, and profit?
 - Will there be entry or exit?
- Suppose that a competitive firm's marginal cost of producing output q is given by $MC(q) = 3 + 2q$. Assume that the market price of its product is \$9:
 - What level of output will the firm produce?
 - What is the firm's producer surplus?
- Suppose that the average variable cost of the firm in Exercise 6 is given by $AVC(q) = 3 + q$. Suppose also that the firm's fixed costs are known to be \$3. Will the firm be earning a positive, negative, or zero profit in the short run?
- A competitive industry is in long-run equilibrium. A sales tax is then placed on all firms. What do you expect to happen to the price of the product, the number of firms in the industry, and the output of each firm?

TABLE 8.A

OUTPUT (UNITS)	PRICE (\$/UNIT)	REVENUE (\$)	TOTAL COST (\$)	PROFIT (\$)	MARGINAL COST (\$)	MARGINAL REVENUE (\$)
0	40	0	50	-50	—	—
1	40	40	100	-60	50	40
2	40	80	128	-48	28	40
3	40	120	148	-28	20	40
4	40	160	162	-2	14	40
5	40	200	180	20	18	40
6	40	240	200	40	20	40
7	40	280	222	58	22	40
8	40	320	260	60	38	40
9	40	360	305	55	45	40
10	40	400	360	40	55	40
11	40	440	425	15	65	40

- A sales tax of 10 percent is placed on half the firms (the polluters) in a competitive industry. The revenue is paid to the remaining firms (the nonpolluters) as a 10 percent subsidy on the value of output sold.
 - Assuming that all firms have identical constant long-run average costs before imposition of the sales tax-subsidy, what do you expect to happen to

the price of the product, the output of each firm, and industry output, both in the short run and the long run? (*Hint*: How does price relate to industry input?)

- Can such a policy *always* be achieved with a balanced budget in which tax revenues are equal to subsidy payments? Why or why not? Explain.

CHAPTER 9

The Analysis of Competitive Markets

In Chapter 2, we saw how supply and demand curves can help us describe and understand the behavior of competitive markets. In Chapters 3 to 8, we saw how these curves are derived and what determines their shapes. Building on this foundation, we return to supply-demand analysis and show how it can be applied to a wide variety of economic problems—problems that might concern a consumer faced with a purchasing decision, a firm faced with a long-range planning problem, or a government agency that has to design a policy and evaluate its likely impact.

We begin by showing how consumer and producer surplus can be used to study the *welfare effects* of a government policy—in other words, who gains and who loses from the policy, and by how much. We also use consumer and producer surplus to demonstrate the *efficiency* of a competitive market—why the equilibrium price and quantity in a competitive market maximizes the aggregate economic welfare of producers and consumers.

Then we apply supply-demand analysis to a variety of problems. Very few markets in the United States have been untouched by government interventions of one kind or another, so most of the problems that we will study deal with the effects of such interventions. Our objective is not simply to solve these problems, but to show you how to use the tools of economic analysis to deal with others like them on your own. We hope that by working through the examples we provide, you will see how to calculate the response of markets to changing economic conditions or government policies and to evaluate the resulting gains and losses to consumers and producers.

Chapter Outline

- 9.1 Evaluating the Gains and Losses from Government Policies—Consumer and Producer Surplus 238
- 9.2 The Efficiency of a Competitive Market 294
- 9.3 Minimum Prices 298
- 9.4 Price Supports and Production Quotas 302
- 9.5 Import Quotas and Tariffs 309
- 9.6 The Impact of a Tax or Subsidy 313

List of Examples

- 9.1 Price Controls and Natural Gas Shortages 292
- 9.2 The Market for Human Kidneys 295
- 9.3 Airline Regulation 300
- 9.4 Supporting the Price of Wheat 306
- 9.5 The Sugar Quota 312
- 9.6 A Tax on Gasoline 318

9.1 Evaluating the Gains and Losses from Government Policies—Consumer and Producer Surplus

We saw at the end of Chapter 2 that a government-imposed price ceiling causes the quantity of a good demanded to rise (at the lower price, consumers want to buy more) and the quantity supplied to fall (producers are not willing to supply as much at the lower price). The result is a shortage—i.e., excess demand. Of course, those consumers who can still buy the good will be better off because they will now pay less. (Presumably, this was the objective of the policy in the first place.) But if we also take into account those who cannot obtain the good, how much better off are consumers *as a whole*? Might they be worse off? And if we lump consumers and producers together, will their *total welfare* be greater or lower, and by how much? To answer questions such as these, we need a way to measure the gains and losses from government interventions and the changes in market price and quantity that such interventions cause.

Our method is to calculate the changes in *consumer and producer surplus* that result from an intervention. In Chapter 4, we saw that *consumer surplus* measures the aggregate net benefit that consumers obtain from a competitive market. In Chapter 8, we saw how *producer surplus* measures the aggregate net benefit to producers. Here we will see how consumer and producer surplus can be applied in practice.

Review of Consumer and Producer Surplus

In an unregulated, competitive market, consumers and producers buy and sell at the prevailing market price. But remember, for some consumers the value of the good *exceeds* this market price; they would pay more for the good if they had to. *Consumer surplus* is the total benefit or value that consumers receive beyond what they pay for the good.

For example, suppose the market price is \$5 per unit, as in Figure 9.1. Some consumers probably value this good very highly and would pay much more than \$5 for it. Consumer A, for example, would pay up to \$10 for the good. However, because the market price is only \$5, he enjoys a net benefit of \$5—the \$10 value he places on the good, less the \$5 he must pay to obtain it. Consumer B values the good somewhat less highly. She would be willing to pay \$7, and thus enjoys a \$2 net benefit. Finally, Consumer C values the good at exactly the market price, \$5. He is indifferent between buying or not buying the good, and if the market price were one cent higher, he would forgo the purchase. Consumer C, therefore, obtains no net benefit.¹

For consumers in the aggregate, consumer surplus is the area between the demand curve and the market price (i.e., the yellow-shaded area in Figure 9.1). Because *consumer surplus measures the total net benefit to consumers*, we can measure the gain or loss to consumers from a government intervention by measuring the resulting change in consumer surplus.

¹ Of course, some consumers value the good at *less* than \$5. These consumers make up the part of the demand curve to the right of the equilibrium quantity Q_0 and will not purchase the good.

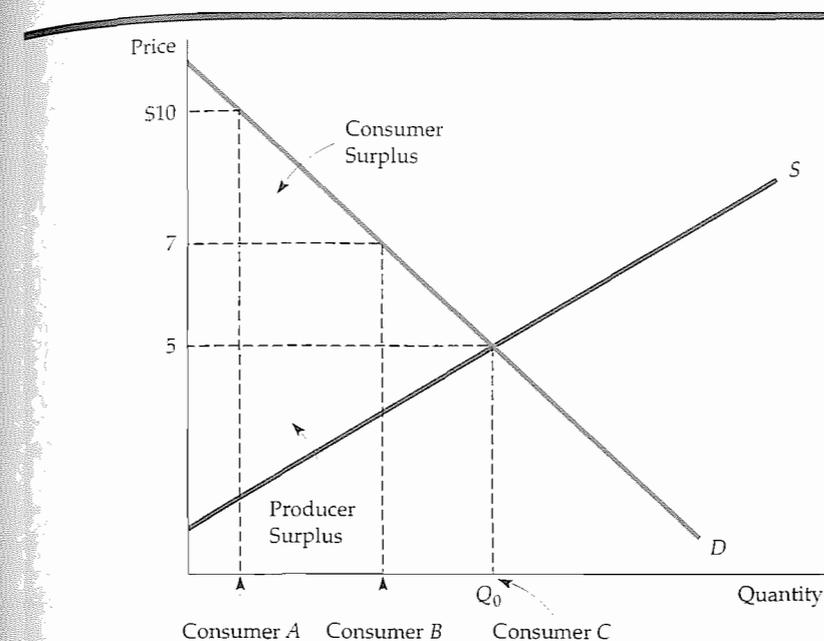


FIGURE 9.1 Consumer and Producer Surplus

Consumer A would pay \$10 for a good whose market price is \$5 and therefore enjoys a benefit of \$5. Consumer B enjoys a benefit of \$2, and Consumer C, who values the good at exactly the market price, enjoys no benefit. Consumer surplus, which measures the total benefit to all consumers, is the yellow-shaded area between the demand curve and the market price. Producer surplus measures the total profits of producers, plus rents to factor inputs. It is the green-shaded area between the supply curve and the market price. Together, consumer and producer surplus measure the welfare benefit of a competitive market.

Producer surplus is the analogous measure for producers. Some producers are producing units at a cost just equal to the market price. Other units, however, could be produced for less than the market price and would still be produced and sold even if the market price were lower. Producers, therefore, enjoy a benefit—a surplus—from selling those units. For each unit, this surplus is the difference between the market price the producer receives and the marginal cost of producing this unit.

For the market as a whole, producer surplus is the area above the supply curve up to the market price; this is *the benefit that lower-cost producers enjoy by selling at the market price*. In Figure 9.1 it is the green triangle. And because producer surplus measures the total net benefit to producers, we can measure the gain or loss to producers from a government intervention by measuring the resulting change in producer surplus.

Application of Consumer and Producer Surplus

With consumer and producer surplus, we can evaluate the *welfare effects* of a government intervention in the market. We can determine who gains and who loses from the intervention, and by how much. To see how this is done, let's return to the example of *price controls* that we first encountered toward the end

For a review of producer surplus, see §8.5, where it is defined as the sum over all units produced of the difference between the market price of the good and the marginal cost of its production.

welfare effects Gains and losses caused by government intervention in the market.

In §2.7, we explain that under price controls, the price of a product can be no higher than a maximum allowable ceiling price.

For a review of consumer surplus, see §4.4, where it is defined as the difference between what a consumer is willing to pay for a good and what the consumer actually pays when buying it.

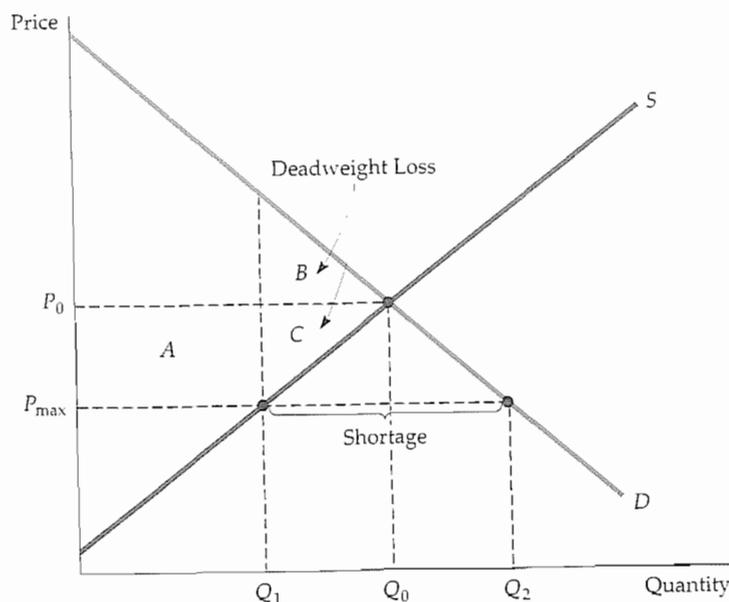


FIGURE 9.2 Change in Consumer and Producer Surplus from Price Controls

The price of a good has been regulated to be no higher than P_{max} , which is below the market-clearing price P_0 . The gain to consumers is the difference between rectangle A and triangle B. The loss to producers is the sum of rectangle A and triangle C. Triangles B and C together measure the deadweight loss from price controls.

of Chapter 2. The government makes it illegal for producers to charge more than a *ceiling price* set below the market-clearing level. Recall that by decreasing production and increasing the quantity demanded, such a price ceiling creates a shortage (excess demand).

Figure 9.2 replicates Figure 2.22, except that it also shows the changes in consumer and producer surplus that result from the government price-control policy. Let's go through these changes step by step.

- 1. Change in Consumer Surplus:** Some consumers are worse off as a result of the policy, and others are better off. The ones who are worse off are those who have been rationed out of the market because of the reduction in production and sales from Q_0 to Q_1 . Other consumers, however, can still purchase the good (perhaps because they are in the right place at the right time, or are willing to wait in line). These consumers are better off because they can buy the good at a lower price (P_{max} rather than P_0).

How much better off or worse off is each group? The consumers who can still buy the good enjoy an *increase* in consumer surplus, which is given by the blue-shaded rectangle A. This rectangle measures the reduction of price in each unit times the number of units consumers are able to buy at the lower price. On the other hand, those consumers who can no longer buy the good lose surplus; their *loss* is given by the green-shaded triangle B. This triangle measures the value to consumers, net of what they would have had to

pay, that is lost because of the reduction in output from Q_0 to Q_1 .² The net change in consumer surplus is therefore $A - B$. In Figure 9.2, because rectangle A is larger than triangle B, we know that the net change in consumer surplus is positive.

- 2. Change in Producer Surplus:** With price controls, some producers (those with relatively lower costs) will stay in the market but will receive a lower price for their output, while other producers will leave the market. Both groups will lose producer surplus. Those producers who remain in the market and produce quantity Q_1 are now receiving a lower price. They have lost the producer surplus given by rectangle A. However, *total* production has also dropped. The purple-shaded triangle C measures the additional loss of producer surplus for those producers who have left the market and those who have stayed in the market but are producing less. Therefore, the total change in producer surplus is $-A - C$. Producers clearly lose as a result of price controls.

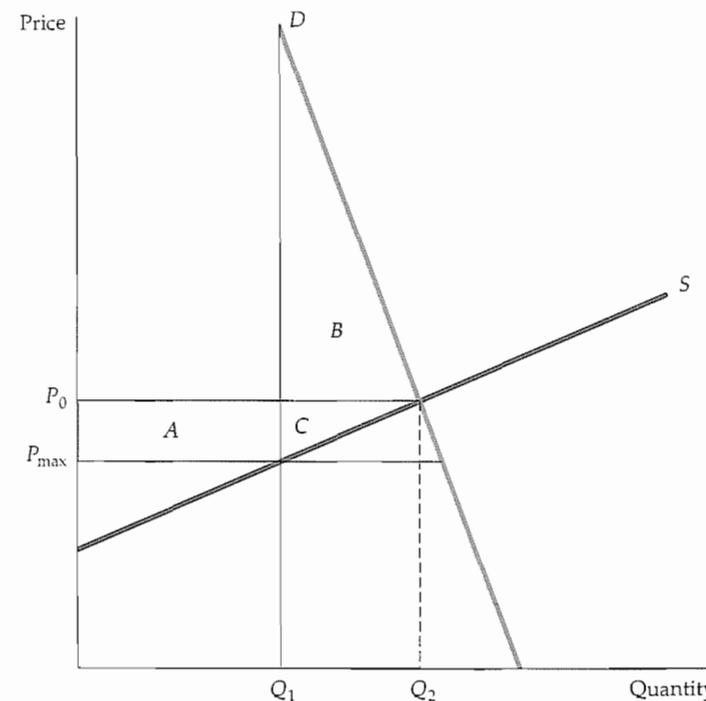


FIGURE 9.3 Effect of Price Controls When Demand Is Inelastic

If demand is sufficiently inelastic, triangle B can be larger than rectangle A. In this case, consumers suffer a net loss from price controls.

² We are assuming here that those consumers who are able to buy the good are the ones that value it most highly. If this were not the case, the amount of lost consumer surplus would be larger than triangle B.

deadweight loss Net loss of total (consumer plus producer) surplus.

3. **Deadweight Loss:** Is the loss to producers from price controls offset by the gain to consumers? No. As Figure 9.2 shows, price controls result in a net loss of total surplus, which we call a **deadweight loss**. Recall that the change in consumer surplus is $A - B$ and the change in producer surplus is $-A - C$. The *total* change in surplus is therefore $(A - B) + (-A - C) = -B - C$. We thus have a deadweight loss, which is given by the two triangles B and C in Figure 9.2. This deadweight loss is an inefficiency caused by price controls; the loss in producer surplus exceeds the gain in consumer surplus.

If politicians value consumer surplus more than producer surplus, this deadweight loss from price controls may not carry much political weight. However, if the demand curve is very inelastic, price controls can result in a *net loss of consumer surplus*, as Figure 9.3 shows. In that figure, triangle B , which measures the loss to consumers who have been rationed out of the market, is larger than rectangle A , which measures the gain to consumers able to buy the good. Here, consumers value the good highly, so those who are rationed out suffer a large loss.

The demand for gasoline is very inelastic in the short run (but much more elastic in the long run). During the summer of 1979, gasoline shortages resulted from oil price controls that prevented domestic gasoline prices from increasing to rising world levels. Consumers spent hours waiting in line to buy gasoline. This was a good example of price controls making consumers—the group whom the policy was presumably intended to protect—worse off.

EXAMPLE 9.1 Price Controls and Natural Gas Shortages

In example 2.9 in Chapter 2, we saw that during the 1970s, price controls created large shortages of natural gas. Today, producers of natural gas, oil, and other fuels are concerned that the government might reimpose controls if prices rise sharply. Therefore it is important to be able to evaluate the welfare effects of price controls. How much did consumers gain from natural gas price controls? How much did producers lose? What was the deadweight loss to the country? We can answer these questions by calculating the resulting changes in consumer and producer surplus.

Basing our analysis on the numbers for 1975, let's calculate the annual gains and losses that resulted from controls. Refer to Example 2.9, where we showed that the supply and demand curves can be approximated as follows:

$$\begin{aligned} \text{Supply: } & Q^S = 14 + 2P_G + 0.25P_O \\ \text{Demand: } & Q^D = -5P_G + 3.75P_O \end{aligned}$$

where Q^S and Q^D are the quantities supplied and demanded, each measured in trillion cubic feet (Tcf), P_G is the price of natural gas in dollars per thousand cubic feet (\$/mcf), and P_O is the price of oil in dollars per barrel (\$/b). As you can verify by setting Q^S equal to Q^D and using a price of oil of \$8 per barrel, the equilibrium free market price and quantity are \$2 per mcf and 20 Tcf, respectively. Under the regulations, however, the maximum allowable price was \$1 per mcf.

Figure 9.4 shows these supply and demand curves and compares the free market and regulated prices. Rectangle A and triangles B and C measure the

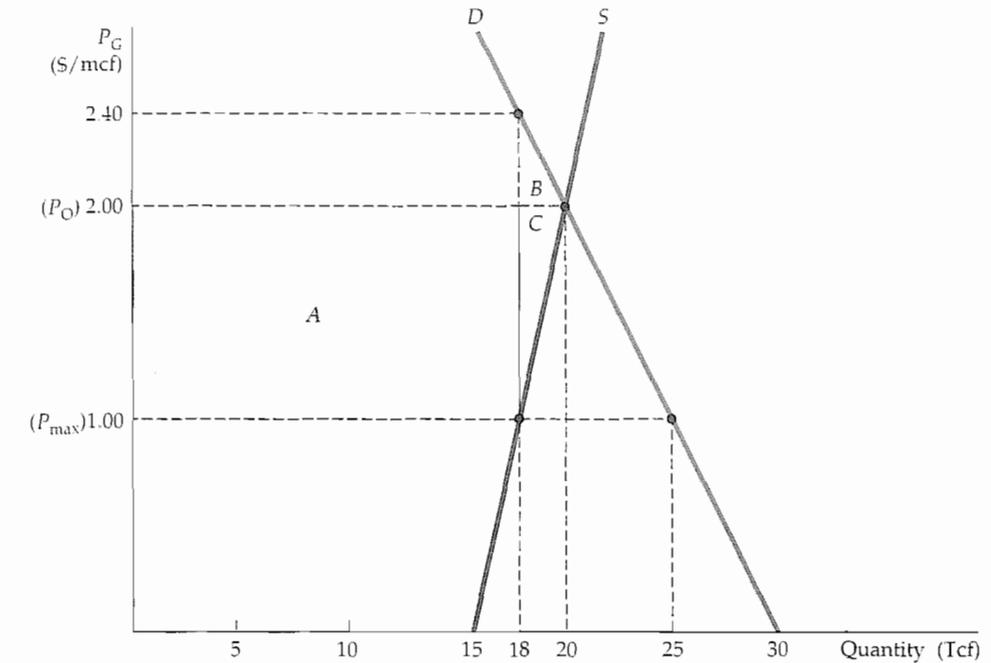


FIGURE 9.4 Effects of Natural Gas Price Controls

The market-clearing price of natural gas is \$2 per mcf, and the maximum allowable price is \$1. A shortage of $25 - 18 = 7$ Tcf results. The gain to consumers is rectangle A minus triangle B , and the loss to producers is rectangle A plus triangle C .

changes in consumer and producer surplus resulting from price controls. By calculating the areas of the rectangle and triangles, we can determine the gains and losses from controls.

To do the calculations, first note that 1 Tcf is equal to 1 billion mcf. (We must put the quantities and prices in common units.) Also, by substituting the quantity 18 Tcf into the equation for the demand curve, we can determine that the vertical line at 18 Tcf intersects the demand curve at a price of \$2.40 per mcf. Then we can calculate the areas as follows:

$$\begin{aligned} A &= (18 \text{ billion mcf}) \times (\$1/\text{mcf}) = \$18 \text{ billion} \\ B &= (1/2) \times (2 \text{ billion mcf}) \times (\$0.40/\text{mcf}) = \$0.4 \text{ billion} \\ C &= (1/2) \times (2 \text{ billion mcf}) \times (\$1/\text{mcf}) = \$1 \text{ billion} \end{aligned}$$

(The area of a triangle is one-half the product of its altitude and its base.)

The 1975 change in consumer surplus resulting from price controls was therefore $A - B = 18 - 0.4 = \$17.6$ billion. The change in producer surplus was $-A - C = -18 - 1 = -\$19$ billion. And finally, the deadweight loss for the year was $-B - C = -0.4 - 1 = -\$1.4$ billion. Remember that this amount of \$1.4 billion per year is in 1975 dollars. In year 2000 dollars, the deadweight loss is more than \$4 billion per year—a significant loss to society.

9.2 The Efficiency of a Competitive Market

economic efficiency Maximization of aggregate consumer and producer surplus.

To evaluate a market outcome, we often ask whether it achieves **economic efficiency**—the maximization of aggregate consumer and producer surplus. We just saw how price controls create a deadweight loss. The policy therefore imposes an *efficiency cost* on the economy: Taken together, producer and consumer surplus are reduced by the amount of the deadweight loss. (Of course, this does not mean that such a policy is bad; it may achieve other objectives that policymakers and the public deem important.)

market failure Situation in which an unregulated competitive market is inefficient because prices fail to provide proper signals to consumers and producers.

Market Failure One might think that if the only objective is to achieve economic efficiency, a competitive market is better left alone. This is sometimes, but not always, the case. In some situations, a **market failure** occurs: Because prices fail to provide the proper signals to consumers and producers, the unregulated competitive market is inefficient—i.e., does not maximize aggregate consumer and producer surplus. There are two important instances in which market failure can occur:

1. **Externalities:** Sometimes the actions of either consumers or producers result in either costs or benefits that do not show up as part of the market price. Such costs or benefits are called **externalities** because they are “external” to the market. One example is the cost to society of environmental pollution by a producer of industrial chemicals. Without government intervention, such a producer will have no incentive to consider the social cost of this pollution. We examine externalities and the proper government response to them in Chapter 18.
2. **Lack of Information:** Market failure can also occur when consumers lack information about the quality or nature of a product and so cannot make utility-maximizing purchasing decisions. Government intervention (e.g., requiring “truth in labeling”) may then be desirable. The role of information is discussed in detail in Chapter 17.

externality Action taken by either a producer or a consumer which affects other producers or consumers but is not accounted for by the market price.

In the absence of externalities or a lack of information, an unregulated competitive market does lead to the economically efficient output level. To see this, let’s consider what happens if price is constrained to be something other than the equilibrium market-clearing price.

We have already examined the effects of a *price ceiling* (a price held below the market-clearing price). As you can see in Figure 9.2, production falls (from Q_0 to Q_1), and there is a corresponding loss of total surplus (the deadweight-loss triangles B and C). Too little is produced, and consumers and producers in the aggregate are worse off.

Now suppose instead that the government required the price to be *above* the market-clearing price—say, P_2 instead of P_0 . As Figure 9.5 shows, although producers would like to produce more at this higher price (Q_2 instead of Q_0), consumers will now buy less (Q_3 instead of Q_0). If we assume that producers produce only what can be sold, the market output level will be Q_3 , and again, there is a net loss of total surplus. In Figure 9.5, rectangle A now represents a transfer from consumers to producers (who now receive a higher price), but triangles B

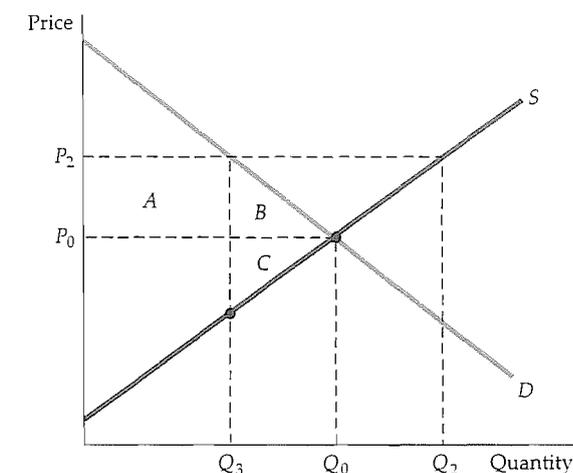


FIGURE 9.5 Welfare Loss When Price Is Held Above Market-Clearing Level

When price is regulated to be no lower than P_2 , only Q_3 will be demanded. If Q_3 is produced, the deadweight loss is given by triangles B and C . At price P_2 , producers would like to produce more than Q_3 . If they do, the deadweight loss will be even larger.

and C are again a deadweight loss. Because of the higher price, some consumers are no longer buying the good (a loss of consumer surplus given by triangle B), and some producers are no longer producing it (a loss of producer surplus given by triangle C).

In fact, the deadweight loss triangles B and C in Figure 9.5 give an optimistic assessment of the efficiency cost of policies that force price above market-clearing levels. Some producers, enticed by the high price P_2 , might increase their capacity and output levels, which would result in unsold output. (This happened in the airline industry when fares were regulated above market-clearing levels by the Civil Aeronautics Board.) Or to satisfy producers, the government might buy up unsold output to maintain production at Q_2 or close to it. (This is what happens in U.S. agriculture.) In both cases, the total welfare loss will exceed triangles B and C .

We will examine minimum prices, price supports, and related policies in some detail in the next few sections. Besides showing how supply-demand analysis can be used to understand and assess these policies, we will see how deviations from the competitive market equilibrium lead to efficiency costs.

EXAMPLE 9.2 The Market for Human Kidneys

Should people have the right to sell parts of their bodies? Congress believes the answer is no. In 1984, it passed the National Organ Transplantation Act, which prohibits the sale of organs for transplantation. Organs may only be donated.

Although the law prohibits their sale, it does not make organs valueless. Instead, it prevents those who supply organs (living persons or the families of the deceased) from reaping their economic value. It also creates a shortage of

organs. Each year, about 8000 kidneys, 20,000 corneas, and 1200 hearts are transplanted in the United States, but there is considerable excess demand for these organs, and many potential recipients must do without them. Some potential recipients die as a result.

To understand the effects of this law, let's consider the supply and demand for kidneys. First the supply curve. Even at a price of zero (the effective price under the law), donors supply about 8000 kidneys per year. But many other people who need kidney transplants cannot obtain them because of a lack of donors. It has been estimated that 4000 more kidneys would be supplied if the price were \$20,000. We can fit a linear supply curve to this data—i.e., a supply curve of the form $Q = a + bP$. When $P = 0$, $Q = 8000$, so $a = 8000$. If $P = \$20,000$, $Q = 12,000$, so $b = (12,000 - 8000)/20,000 = 0.2$. Thus the supply curve is

$$\text{Supply: } Q^S = 8000 + 0.2P$$

Note that at a price of \$20,000, the elasticity of supply is 0.33.

It is expected that at a price of \$20,000, the number of kidneys demanded would be 12,000 per year. Like supply, demand is relatively price inelastic; a reasonable estimate for the elasticity of demand at the \$20,000 price is -0.33 . This implies the following linear demand curve:

$$\text{Demand: } Q^D = 16,000 - 0.2P$$

These supply and demand curves are plotted in Figure 9.6, which shows the market-clearing price and quantity of \$20,000 and 12,000, respectively.

Because the sale of kidneys is prohibited, supply is limited to 8000 (the number of kidneys that people donate). This constrained supply is shown as the vertical line S' . How does this affect the welfare of kidney suppliers and recipients?

First consider suppliers. Those who provide kidneys fail to receive the \$20,000 that each kidney is worth—a loss of surplus represented by rectangle A and equal to $(8000)(\$20,000) = \160 million. Moreover, some people who would supply kidneys if they were paid do not. These people lose an amount of surplus represented by triangle C , which is equal to $(1/2)(4000)(\$20,000) = \40 million. Therefore the total loss to suppliers is \$200 million.

What about recipients? Presumably the law intended to treat the kidney as a gift to the recipient. In this case, those recipients who obtain kidneys gain rectangle A (\$160 million) because they do not have to pay the \$20,000 price. Those who cannot obtain kidneys lose surplus of an amount given by triangle B and equal to \$40 million. This implies a net increase in the surplus of recipients of $\$160 - \$40 = \$120$ million. It also implies a deadweight loss equal to the areas of triangles B and C (i.e., \$80 million).

These estimates of the welfare effects of the policy may need adjustment for two reasons. First, kidneys will not necessarily be allocated to those who value them most highly. If the limited supply of kidneys is partly allocated to people with valuations below \$40,000, the true deadweight loss will be higher than our estimate. Second, with excess demand, there is no way to ensure that recipients will receive their kidneys as gifts. In practice, kidneys are often rationed on the basis of willingness to pay, and many recipients end up paying all or most of the \$40,000 price that is needed to clear the market when supply is constrained

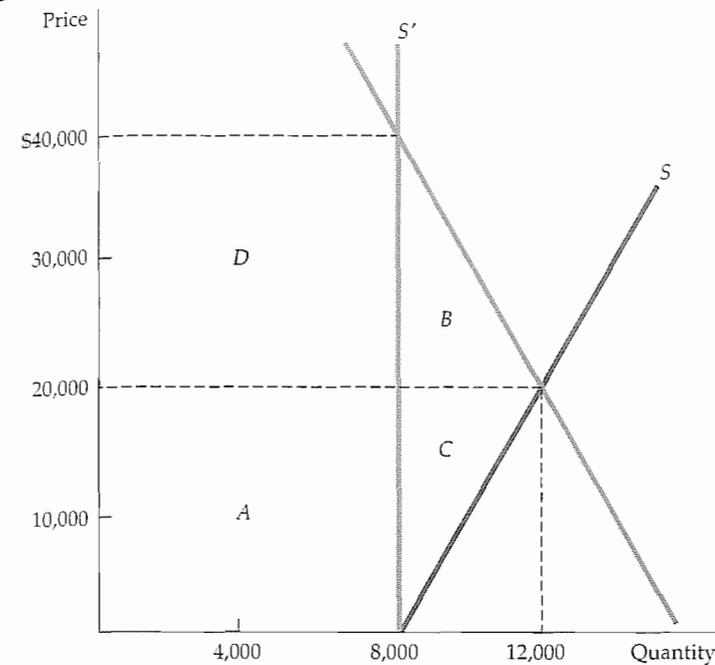


FIGURE 9.6 The Market for Kidneys and the Effect of the National Organ Transplantation Act

The market-clearing price is \$20,000; at this price, about 12,000 kidneys per year would be supplied. The act effectively makes the price zero. About 8000 kidneys per year are still donated; this constrained supply is shown as S' . The loss to suppliers is given by rectangle A and triangle C . If consumers received kidneys at no cost, their gain would be given by rectangle A less triangle B . In practice, kidneys are often rationed on the basis of willingness to pay, and many recipients pay most or all of the \$40,000 price that clears the market when supply is constrained. Rectangles A and D measure the total value of kidneys when supply is constrained.

to 8000. A good part of the value of the kidneys—rectangles A and D in the figure—is then captured by hospitals and middlemen. As a result, the law reduces the surplus of recipients as well as of suppliers.³

There are, of course, arguments in favor of prohibiting the sale of organs.⁴ One argument stems from the problem of imperfect information; if people receive payment for organs, they may hide adverse information about their health histories. This argument is probably most applicable to the sale of blood, where there is a possibility of transmitting hepatitis, AIDS, or other viruses. But

³ For further analyses of these efficiency costs, see Dwane L. Barney and R. Larry Reynolds, "An Economic Analysis of Transplant Organs," *Atlantic Economic Journal* 17 (September 1989): 12–20; David L. Kaserman and A. H. Barnett, "An Economic Analysis of Transplant Organs: A Comment and Extension," *Atlantic Economic Journal* 19 (June 1991): 57–64; and A. Frank Adams III, A. H. Barnett, and David L. Kaserman, "Markets for Organs: The Question of Supply," *Contemporary Economic Policy*, 17 (April 1999): 147–55.

⁴ For discussions of the strengths and weaknesses of these arguments, see Susan Rose-Ackerman, "Inalienability and the Theory of Property Rights," *Columbia Law Review* 85 (June 1985): 931–69, and Roger D. Blair and David L. Kaserman, "The Economics and Ethics of Alternative Cadaveric Organ Procurement Policies," *Yale Journal on Regulation* 8 (Summer 1991): 403–52.

In §2.6, we explain how to fit linear demand and supply curves from information about the equilibrium price and quantity and the price elasticities of demand and supply.

even here, screening (at a cost that would be included in the market price) may be more efficient than prohibiting sales. This issue has been central to the debate in the United States over blood policy.

A second argument is that it is simply unfair to allocate a basic necessity of life on the basis of ability to pay. This argument transcends economics. However, two points should be kept in mind. First, when the price of a good that has a significant opportunity cost is forced to zero, there is bound to be reduced supply and excess demand. Second, it is not clear why live organs should be treated differently from close substitutes; artificial limbs, joints, and heart valves, for example, are sold even though real kidneys are not.

Many complex ethical and economic issues are involved in the sale of organs. These issues are important, and this example is not intended to sweep them away. Economics, the dismal science, simply shows us that human organs have economic value that cannot be ignored, and that prohibiting their sale imposes a cost on society that must be weighed against the benefits.

9.3 Minimum Prices

As we have seen, government policy sometimes seeks to raise prices above market-clearing levels, rather than lower them. Examples include the former regulation of the airlines by the Civil Aeronautics Board, the minimum wage law, and a variety of agricultural policies. (Most import quotas and tariffs also have this intent, as we will see in Section 9.5.) One way to raise prices above market-clearing levels is by direct regulation—simply make it illegal to charge a price lower than a specific minimum level.

Look again at Figure 9.5. If producers correctly anticipate that they can sell only the lower quantity Q_3 , the net welfare loss will be given by triangles B and C . But as we explained, producers might not limit their output to Q_3 . What happens if producers think they can sell all they want at the higher price and produce accordingly?

This situation is illustrated in Figure 9.7, where P_{\min} denotes a minimum price set by the government. The quantity supplied is now Q_2 and the quantity demanded is Q_3 , the difference representing excess, unsold supply. Now let us follow the resulting changes in consumer and producer surplus.

Those consumers who still purchase the good must now pay a higher price and so suffer a loss of surplus, which is given by rectangle A in Figure 9.7. Some consumers have also dropped out of the market because of the higher price, with a corresponding loss of surplus given by triangle B . The total change in consumer surplus is therefore

$$\Delta CS = -A - B$$

Consumers clearly are worse off as a result of this policy.

What about producers? They receive a higher price for the units they sell, which results in an increase of surplus, given by rectangle A . (Rectangle A represents a transfer of money from consumers to producers). But the drop in sales from Q_0 to Q_3 results in a loss of surplus, which is given by triangle C . Finally, consider the cost to producers of expanding production from Q_0 to Q_2 . Because they sell only Q_3 , there is no revenue to cover the cost of producing $Q_2 - Q_3$.

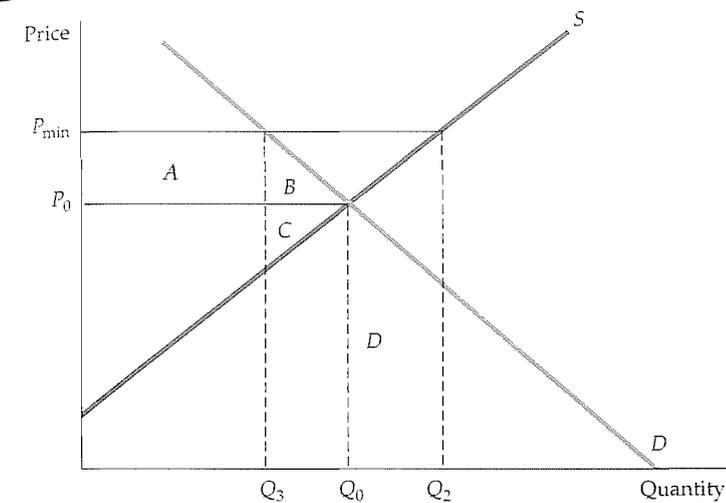


FIGURE 9.7 Price Minimum

Price is regulated to be no lower than P_{\min} . Producers would like to supply Q_2 , but consumers will buy only Q_3 . If producers indeed produce Q_2 , the amount $Q_2 - Q_3$ will go unsold and the change in producer surplus will be $A - C - D$. In this case, producers as a group may be worse off.

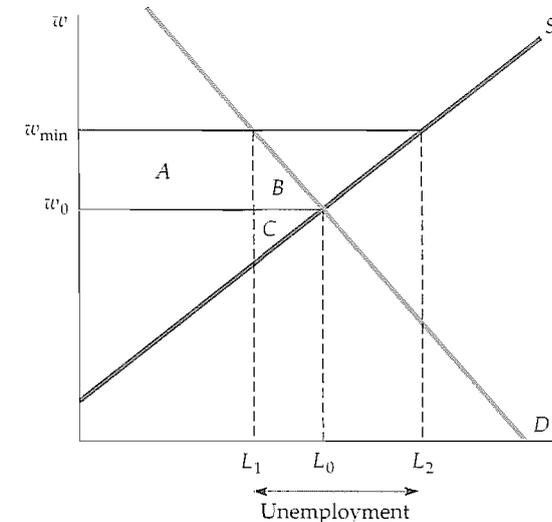


FIGURE 9.8 The Minimum Wage

Although the market-clearing wage is w_0 , firms are not allowed to pay less than w_{\min} . This results in unemployment of an amount $L_2 - L_1$ and a deadweight loss given by triangles B and C .

How can we measure this cost? Remember that the supply curve is the aggregate marginal cost curve for the industry. The supply curve therefore gives us the additional cost of producing each incremental unit. Thus the area under the supply curve from Q_3 to Q_2 is the cost of producing the quantity $Q_2 - Q_3$. This cost is represented by the shaded trapezoid D . So unless producers respond to unsold output by cutting production, the total change in producer surplus is

$$\Delta PS = A - C - D$$

Given that trapezoid D can be large, a minimum price can even result in a net loss of surplus to producers alone! As a result, this form of government intervention can reduce producers' profits because of the cost of excess production.

Another example of a government-imposed price minimum is the minimum wage law. The effect of this policy is illustrated in Figure 9.8, which shows the supply and demand for labor. The wage is set at w_{\min} , a level higher than the market-clearing wage w_0 . As a result, those workers who can find jobs obtain a higher wage. However, some people who want to work will be unable to. The policy results in unemployment, which in the figure is $L_2 - L_1$. We will examine the minimum wage in more detail in Chapter 14.

EXAMPLE 9.3 Airline Regulation

Before 1980, the airline industry in the United States looked very different than it does today. Fares and routes were tightly regulated by the Civil Aeronautics Board (CAB). The CAB set most fares well above what would have prevailed in a free market. It also restricted entry, so that many routes were served by only one or two airlines. By the late 1970s, however, the CAB liberalized fare regulation and allowed airlines to serve any routes they wished. By 1981, the industry had been completely deregulated, and the CAB itself was dissolved in 1982. Since that time, many new airlines have begun service, and price competition is often intense.

Many airline executives feared that deregulation would lead to chaos in the industry, with competitive pressure causing sharply reduced profits and even bankruptcies. After all, the original rationale for CAB regulation was to provide "stability" in an industry that was considered vital to the U.S. economy. And one might think that as long as price was held above its market-clearing level, profits would be higher than they would be in a free market.

Deregulation did lead to major changes in the industry. Some airlines merged or went out of business as new ones entered the industry. Although prices fell considerably (to the benefit of consumers), profits overall did not fall much because the CAB's minimum prices had caused inefficiencies and artificially high costs. The effect of minimum prices is illustrated in Figure 9.9, where P_0 and Q_0 are the market-clearing price and quantity, P_{\min} is the minimum price set by the CAB, and Q_1 is the amount demanded at this higher price. The problem was that at price P_{\min} , airlines wanted to supply a quantity Q_2 , much larger than Q_1 . Although they did not expand output to Q_2 , they did expand it well beyond Q_1 —to Q_3 in the figure—hoping to sell this quantity at the expense of competitors. As a result, load factors (the percentage of seats filled) were relatively low, and so were profits. (Trapezoid D measures the cost of unsold output.)

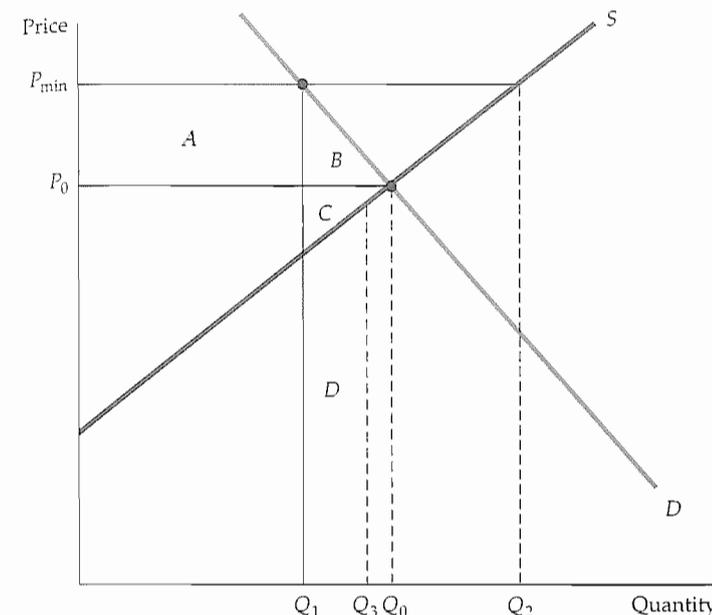


FIGURE 9.9 Effect of Airline Regulation by the Civil Aeronautics Board

At price P_{\min} , airlines would like to supply Q_2 , well above the quantity Q_1 that consumers will buy. Here they supply Q_3 . Trapezoid D is the cost of unsold output. Airline profits may have been lower as a result of regulation because triangle C and trapezoid D can together exceed rectangle A . In addition, consumers lose $A + B$.

Table 9.1 gives some key numbers that illustrate the evolution of the industry.⁵ The number of carriers increased dramatically after deregulation, as did passenger load factors. The passenger-mile rate (the revenue per passenger-mile flown) fell sharply in real (inflation-adjusted) terms from 1980 to 1985, and then continued to drop from 1985 to 1996. This decline was the result of increased competition and reductions in fares. And what about costs? The real cost index indicates that even after adjusting for inflation, costs increased by about 20 percent from 1975 to 1980. But this was largely due to the sharp

TABLE 9.1 Airline Industry Data

	1975	1980	1985	1990	1995	1996
Number of carriers	33	72	86	60	86	96
Passenger load factor (%)	54	59	61	62	67	69
Passenger-mile rate (constant 1995 dollars)	.218	.210	.166	.150	.129	.126
Real cost index (1995 = 100)	101	122	111	107	100	99
Real cost index corrected for fuel increases	94	98	98	100	100	98

⁵ Department of Commerce, *U.S. Statistical Abstract*, 1986, 1989, 1992, 1995, 1998.

increase in fuel costs (caused by the increase in oil prices) that occurred during this period, and had nothing to do with deregulation. The last line in Table 9.1 is the real cost index after adjusting for fuel cost increases. This is what costs would have been had oil prices increased only at the rate of inflation. This index rose only slightly.

What, then, did airline deregulation do for consumers and producers? As new airlines entered the industry and fares went down, consumers benefited. This fact is borne out by the increase in consumer surplus given by rectangle *A* and triangle *B* in Figure 9.9. (The actual benefit to consumers was somewhat smaller than this because *quality* declined as planes became more crowded and delays and cancellations multiplied.) As for the airlines, they had to learn to live in a more competitive—and therefore more turbulent—environment, and some firms did not survive. But overall, airlines became so much more cost-efficient that producer surplus may have increased. The total welfare gain from deregulation was positive, and quite large.⁶

9.4 Price Supports and Production Quotas

Besides imposing a minimum price, the government can increase the price of a good in other ways. Much of American agricultural policy is based on a system of **price supports**, whereby the government sets the market price of a good above the free-market level and buys up whatever output is needed to maintain that price. The government can also increase prices by *restricting production*, either directly or through incentives to producers. In this section, we show how these policies work and examine their impact on consumers, producers, and the federal budget.

Price Supports

In the United States, price supports aim to increase the prices of dairy products, tobacco, corn, peanuts, and so on, so that the producers of those goods can receive higher incomes. Under a price support program, the government sets a support price P_s and then buys up whatever output is needed to keep the market price at this level. Figure 9.10 illustrates this. Let's examine the resulting gains and losses to consumers, producers, and the government.

Consumers At price P_s , the quantity consumers demand falls to Q_1 , but the quantity supplied increases to Q_2 . To maintain this price and avoid having inventories pile up in producer warehouses, the government must buy the quantity $Q_s = Q_2 - Q_1$. In effect, the government adds its demand Q_s to the demand of consumers, and producers can sell all they want at price P_s .

⁶ Studies of the effects of deregulation include John M. Trapani and C. Vincent Olson, "An Analysis of the Impact of Open Entry on Price and the Quality of Service in the Airline Industry," *Review of Economics and Statistics* 64 (February 1982): 118–38; David R. Graham, Daniel P. Kaplan, and David S. Sibley, "Efficiency and Competition in the Airline Industry," *Bell Journal of Economics* (Spring 1983): 118–38; S. Morrison and Clifford Whinston, *The Economic Effects of Airline Deregulation* (Washington: Brookings Institution, 1986); and Nancy L. Rose, "Profitability and Product Quality: Economic Determinants of Airline Safety Performance," *Journal of Political Economy* 98 (October 1990): 944–64.

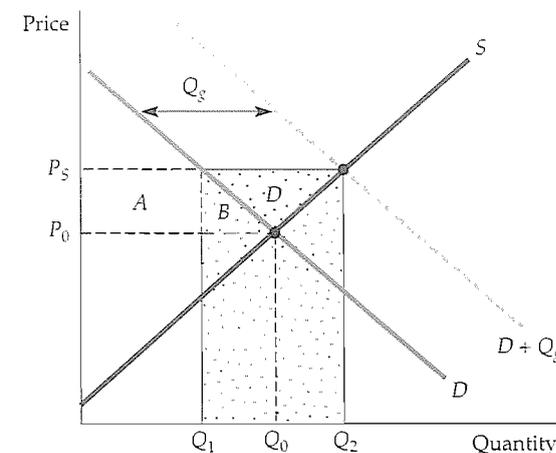


FIGURE 9.10 Price Supports

To maintain a price P_s above the market-clearing price P_0 , the government buys a quantity Q_s . The gain to producers is $A + B + D$. The loss to consumers is $A + B$. The cost to the government is the speckled rectangle, the area of which is $P_s(Q_2 - Q_1)$.

Those consumers who purchase the good must pay the higher price P_s instead of P_0 , so they suffer a loss of consumer surplus given by rectangle *A*. Because of the higher price, other consumers no longer buy the good or buy less of it, and their loss of surplus is given by triangle *B*. So as with the minimum price that we examined above, consumers lose, in this case by an amount

$$\Delta CS = -A - B$$

Producers On the other hand, producers gain (which is why such a policy is implemented). Producers are now selling a larger quantity Q_2 instead of Q_0 , and at a higher price P_s . Observe from Figure 9.10 that producer surplus increases by the amount

$$\Delta PS = A + B + D$$

The Government But there is also a cost to the government (which must be paid for by taxes, and so is ultimately a cost to consumers). That cost is $(Q_2 - Q_1)P_s$, which is what the government must pay for the output it purchases. In Figure 9.10, this is the large speckled rectangle. This cost may be reduced if the government can "dump" some of its purchases—i.e., sell them abroad at a low price. Doing so, however, hurts the ability of domestic producers to sell in foreign markets, and it is domestic producers that the government is trying to please in the first place.

What is the total welfare cost of this policy? To find out, we add the change in consumer surplus to the change in producer surplus and then subtract the cost to the government. Thus the total change in welfare is

$$\Delta CS + \Delta PS - \text{Cost to Govt.} = D - (Q_2 - Q_1)P_s$$

In terms of Figure 9.10, society as a whole is worse off by an amount given by the large speckled rectangle, less triangle *D*.

As we will see in Example 9.4, this welfare loss can be very large. But the most unfortunate part of this policy is the fact that there is a much more efficient way to help farmers. If the objective is to give farmers an additional income equal to $A + B + D$, it is far less costly to society to give them this money directly rather than via price supports. Since consumers are losing $A + B$ anyway with price supports, by paying farmers directly, society saves the large speckled rectangle, less triangle D . Then why doesn't the government simply give farmers money? Perhaps because price supports are a less obvious giveaway and, therefore, politically more attractive.⁷

Production Quotas

Besides entering the market and buying up output—thereby increasing total demand—the government can also cause the price of a good to rise by *reducing supply*. It can do this by decree—that is, by simply setting quotas on how much each firm can produce. With appropriate quotas, the price can then be forced up to any arbitrary level.

This is exactly how many city governments maintain high taxi fares. They limit total supply by requiring each taxicab to have a medallion, and then limit the total number of medallions.⁸ Another example is the control of liquor licenses by state governments. By requiring any bar or restaurant that serves alcohol to have a liquor license and then by limiting the number of licenses, entry by new restaurateurs is limited, which allows those who have the licenses to earn higher prices and profit margins.

The welfare effects of production quotas are shown in Figure 9.11. The government restricts the quantity supplied to Q_1 , rather than the market-clearing level Q_0 . Thus the supply curve becomes the vertical line S' at Q_1 . Consumer surplus is reduced by rectangle A (those consumers who buy the good pay a higher price) plus triangle B (at this higher price, some consumers no longer purchase the good). Producers gain rectangle A (by selling at a higher price) but lose triangle C (because they now produce and sell Q_1 rather than Q_0). Once again, there is a deadweight loss, given by triangles B and C .

Incentive Programs In U.S. agricultural policy, output is reduced by incentives rather than by outright quotas. *Acreage limitation programs* give farmers financial incentives to leave some of their acreage idle. Figure 9.11 also shows the welfare effects of reducing supply in this way. Note that because farmers agree to limit the acreage planted, the supply curve again becomes completely inelastic at the quantity Q_1 , and the market price is increased from P_0 to P_s .

As with direct production quotas, the change in consumer surplus is

$$\Delta CS = -A - B$$

⁷ In practice, price supports for many agricultural commodities are effected through loans. The loan rate is in effect a price floor. If during the loan period market prices are not sufficiently high, farmers can forfeit their grain to the government (specifically to the Commodity Credit Corporation) as *full payment for the loan*. Farmers have the incentive to do this unless the market price rises above the support price.

⁸ For example, as of 1995 New York City had not issued any new taxi medallions for half a century. Only 11,800 taxis were permitted to cruise the city's streets, the same number as in 1937! As a result, in 1995 a medallion could be sold for about \$120,000. It shouldn't be a surprise, then, that the city's taxicab companies have vigorously opposed phasing out medallions in favor of an open system. Washington, D.C., has such an open system: An average taxi ride there costs about half of what it does in New York, and taxis are far more available.

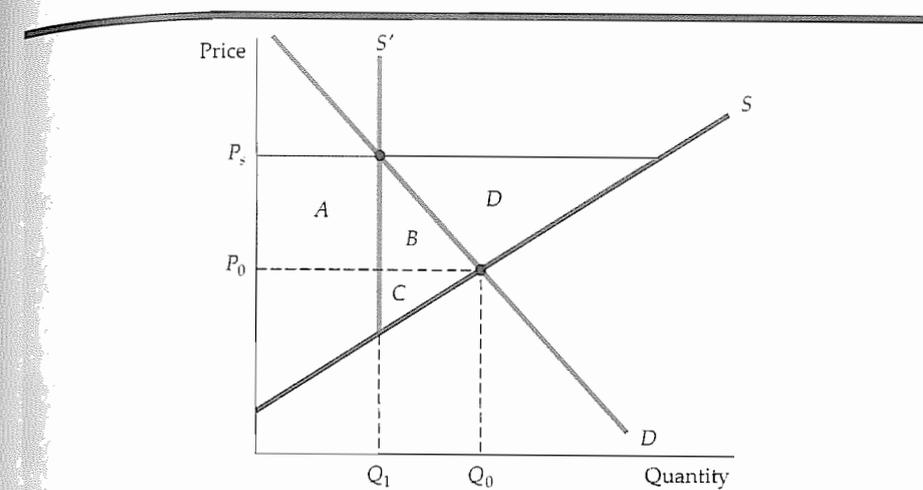


FIGURE 9.11 Supply Restrictions

To maintain a price P_s above the market-clearing price P_0 , the government can restrict supply to Q_1 , either by imposing production quotas (as with taxi cab medallions) or by giving producers a financial incentive to reduce output (as with acreage limitations). For an incentive to work, it must be at least as large as $B + C + D$, which is the additional profit earned by planting, given the higher price P_s . The cost to the government is therefore at least $B + C + D$.

Farmers now receive a higher price for the production Q_1 , which corresponds to a gain in surplus of rectangle A . But because production is reduced from Q_0 to Q_1 , there is a loss of producer surplus corresponding to triangle C . Finally, farmers receive money from the government as an incentive to reduce production. Thus, the total change in producer surplus is now

$$\Delta PS = A - C + \text{Payments for not producing}$$

The cost to the government is a payment sufficient to give farmers an incentive to reduce output to Q_1 . That incentive must be at least as large as $B + C + D$ because that is the additional profit that could be made by planting, given the higher price P_s . (Remember that the higher price P_s gives farmers an incentive to produce *more* even though the government is trying to get them to produce *less*.) Thus the cost to the government is at least $B + C + D$, and the total change in producer surplus is

$$\Delta PS = A - C + B + C + D = A + B + D$$

This is the same change in producer surplus as with price supports maintained by government purchases of output. (Refer to Figure 9.10.) Farmers, then, should be indifferent between the two policies because they end up gaining the same amount of money from each. Likewise, consumers lose the same amount of money.

Which policy costs the government more? The answer depends on whether the sum of triangles $B + C + D$ in Figure 9.11 is larger or smaller than $(Q_2 - Q_1)P_s$ (the large speckled rectangle) in Figure 9.10. Usually it will be smaller, so that an acreage limitation program costs the government (and society) less than price supports maintained by government purchases.

Still, even an acreage limitation program is more costly to society than simply handing the farmers money. The total change in welfare ($\Delta CS + \Delta PS - \text{Cost to Govt.}$) under the acreage limitation program is

$$\Delta \text{Welfare} = -A - B + A + B + D - B - C - D = -B - C$$

Society would clearly be better off in efficiency terms if the government simply gave the farmers $A + B + D$, leaving price and output alone. Farmers would then gain $A + B + D$ and the government would lose $A + B + D$, for a total welfare change of zero, instead of a loss of $B + C$. However, economic efficiency is not always the objective of government policy.

EXAMPLE 9.4 Supporting the Price of Wheat

In Examples 2.4 and 4.3, we began to examine the market for wheat in the United States. Using linear demand and supply curves, we found that the market-clearing price of wheat was about \$3.46 in 1981, but it fell to about \$2.65 by 1998 because of a drop in export demand. In fact, government programs kept the actual price of wheat higher and provided direct subsidies to farmers. How did these programs work, how much did they end up costing consumers, and how much did they add to the federal deficit?

First, let us examine the market in 1981. In that year there were no effective limitations on the production of wheat, but price was increased to \$3.70 by government purchases. How much would the government have had to buy to get the price from \$3.46 to \$3.70? To answer this, first write the equations for supply, and for total (domestic plus export) demand:

$$1981 \text{ Supply: } Q_S = 1800 + 240P$$

$$1981 \text{ Demand: } Q_D = 3550 - 266P$$

By equating supply and demand, you can check that the market-clearing price is \$3.46, and that the quantity produced is 2630 million bushels. Figure 9.12 illustrates this.

To increase the price to \$3.70, the government must buy a quantity of wheat Q_g . Total demand (private plus government) will then be

$$1981 \text{ Total demand: } Q_{DT} = 3550 - 266P + Q_g$$

Now equate supply with this total demand:

$$1800 + 240P = 3550 - 266P + Q_g$$

or

$$Q_g = 506P - 1750$$

This equation can be used to determine the required quantity of government wheat purchases Q_g as a function of the desired support price P . To achieve a price of \$3.70, the government must buy

$$Q_g = (506)(3.70) - 1750 = 122 \text{ million bushels}$$

Note in Figure 9.12 that these 122 million bushels are the difference between the quantity supplied at the \$3.70 price (2688 million bushels) and the quantity of private demand (2566 million bushels). The figure also shows the gains and losses to consumers and producers. Recall that consumers lose rectangle A and

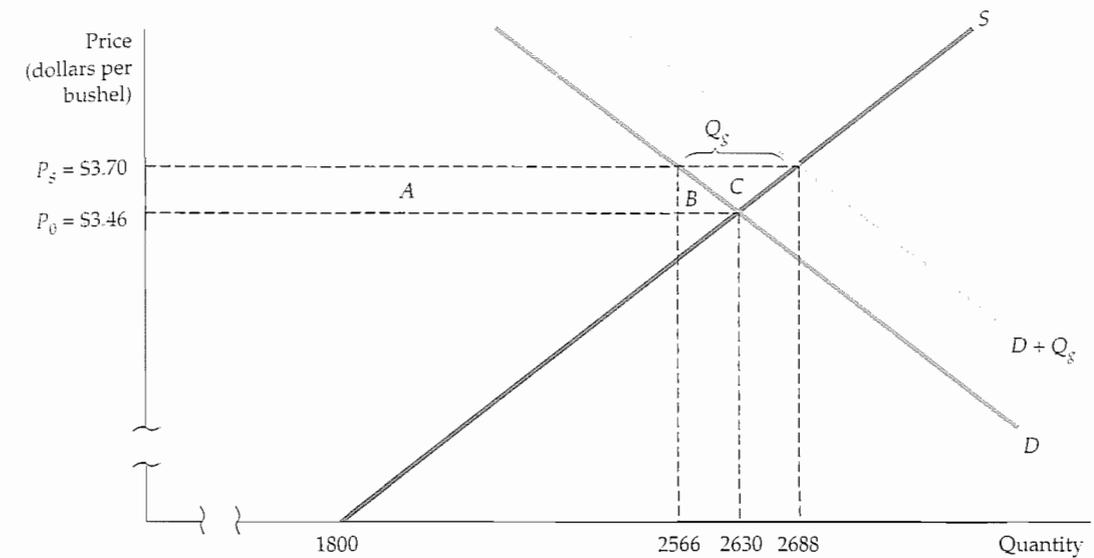


FIGURE 9.12 The Wheat Market in 1981

By buying 122 million bushels of wheat, the government increased the market-clearing price from \$3.46 per bushel to \$3.70.

triangle B . You can verify that rectangle A is $(3.70 - 3.46)(2566) = \$616$ million, and triangle B is $(1/2)(3.70 - 3.46)(2630 - 2566) = \8 million, so the total cost to consumers is \$624 million.

The cost to the government is the \$3.70 it pays for the wheat times the 122 million bushels it buys, or \$451.4 million. The total cost of the program is then $\$624 + \$451.4 = \$1075$ million. Compare this with the gain to producers, which is rectangle A plus triangles B and C . You can verify that this gain is \$638 million.

Price supports for wheat were expensive in 1981. To increase the surplus of farmers by \$638 million, consumers and taxpayers had to pay \$1076 million. In fact taxpayers paid even more. Wheat producers were also given subsidies of about 30 cents per bushel, which adds up to another \$806 million.

In 1985 the situation became even worse because of the drop in export demand. In that year the supply and demand curves were as follows:

$$1985 \text{ Supply: } Q_S = 1800 + 240P$$

$$1985 \text{ Demand: } Q_D = 2580 - 194P$$

You can verify that the market-clearing price and quantity were \$1.80 and 2231 million bushels, respectively. The actual price, however, was \$3.20.

To increase the price to \$3.20, the government bought wheat and imposed a production quota of about 2425 million bushels. (Farmers who wanted to take part in the subsidy program—and most did—had to agree to limit their acreage.) Figure 9.13 illustrates this situation. At the quantity 2425 million bushels, the supply curve becomes vertical. Now to determine how much wheat Q_g the government had to buy, set this quantity of 2425 equal to total demand:

$$2425 = 2580 - 194P + Q_g$$

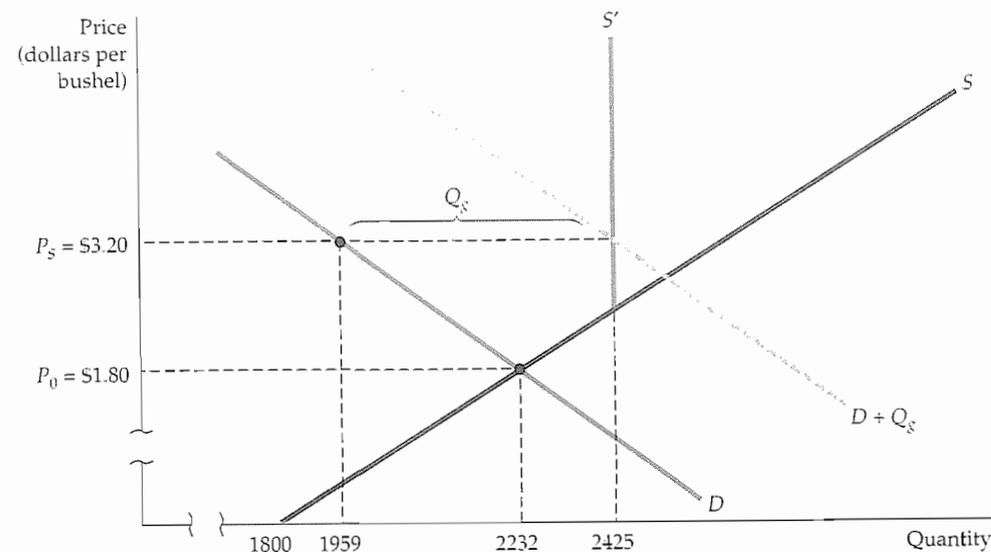


FIGURE 9.13 The Wheat Market in 1985

In 1985 the demand for wheat was much lower than in 1981, so the market-clearing price was only \$1.80. To increase the price to \$3.20, the government bought 466 million bushels and also imposed a production quota of 2425 million bushels.

or

$$Q_s = -155 + 194P$$

Substituting \$3.20 for P , we see that Q_s must be 466 million bushels. This cost the government $(\$3.20)(466) = \1491 million.

Again, this is not the whole story. The government also provided a subsidy of 80 cents per bushel, so that producers again received about \$4.00 for their wheat. Since 2425 million bushels were produced, that subsidy cost an additional \$1940 million. In all, U.S. wheat programs cost taxpayers nearly \$3.5 billion in 1985. Of course, there was also a loss of consumer surplus and a gain of producer surplus; you can calculate what they were.

In 1996, the U.S. Congress passed a new farm bill, nicknamed the “Freedom to Farm” law. It is designed to reduce the role of government and to make agriculture more market oriented. The law eliminates production quotas (for wheat, corn, rice, and other products) and gradually reduces government purchases and subsidies through 2003. However, the law does not completely deregulate U.S. agriculture. For example, price support programs for peanuts and sugar will remain in place. Furthermore, unless Congress renews the law in 2003, pre-1996 price supports and production quotas will go back into effect. Even under the new law, agricultural subsidies remain substantial.

In Example 2.4, we saw that the market-clearing price of wheat in 1998 had dropped to \$2.65 per bushel. The supply and demand curves in 1998 were as follows:

$$\text{Demand: } Q_D = 3244 - 283P$$

$$\text{Supply: } Q_S = 1944 + 207P$$

You can check to see that the market-clearing quantity is 2493 million bushels. Although the government did not buy any wheat in 1998, it provided a direct subsidy to farmers of 66 cents per bushel. Thus the total cost to taxpayers of this subsidy was more than \$1.6 billion.

In 1999, Congress expanded subsidies for wheat, soybeans, and corn by passing an “emergency” agricultural aid bill. The direct cost to taxpayers of these subsidies was estimated at \$24 billion, and this sum is expected to grow in the year 2000 and beyond.⁹

9.5 Import Quotas and Tariffs

Many countries use import quotas and tariffs to keep the domestic price of a product above world levels and thereby enable the domestic industry to enjoy higher profits than it would under free trade. As we will see, the cost to society from this protection can be high, with the loss to consumers exceeding the gain to domestic producers.

Without a quota or tariff, a country will import a good when its world price is below the market price that would prevail if there were no imports. Figure 9.14 shows this. S and D are the domestic supply and demand curves. If there were no imports, the domestic price and quantity would be P_0 and Q_0 , which equate supply and demand. But the world price P_w is below P_0 , so domestic consumers have an incentive to purchase from abroad and will do so if imports are not restricted. How much will be imported? The domestic price will fall to the world price P_w ; at this lower price, domestic production will fall to Q_s , and domestic consumption will rise to Q_d . Imports are then the difference between domestic consumption and domestic production, $Q_d - Q_s$.

Now suppose the government, bowing to pressure from the domestic industry, eliminates imports by imposing a quota of zero—that is, forbidding any importation of the good. What are the gains and losses from such a policy?

With no imports allowed, the domestic price will rise to P_0 . Consumers who still purchase the good (in quantity Q_0) will pay more and will lose an amount of surplus given by trapezoid A and triangle B . Also, given this higher price, some consumers will no longer buy the good, so there is an additional loss of consumer surplus, given by triangle C . The total change in consumer surplus is therefore

$$\Delta CS = -A - B - C$$

What about producers? Output is now higher (Q_0 instead of Q_s) and is sold at a higher price (P_0 instead of P_w). Producer surplus therefore increases by the amount of trapezoid A :

$$\Delta PS = A$$

The change in total surplus, $\Delta CS + \Delta PS$, is therefore $-B - C$. Again, there is a deadweight loss—consumers lose more than producers gain.

import quota Limit on the quantity of a good that can be imported.

tariff Tax on an imported good.

⁹ “It’s Raining Farm Subsidies,” *New York Times*, August 8, 1999.

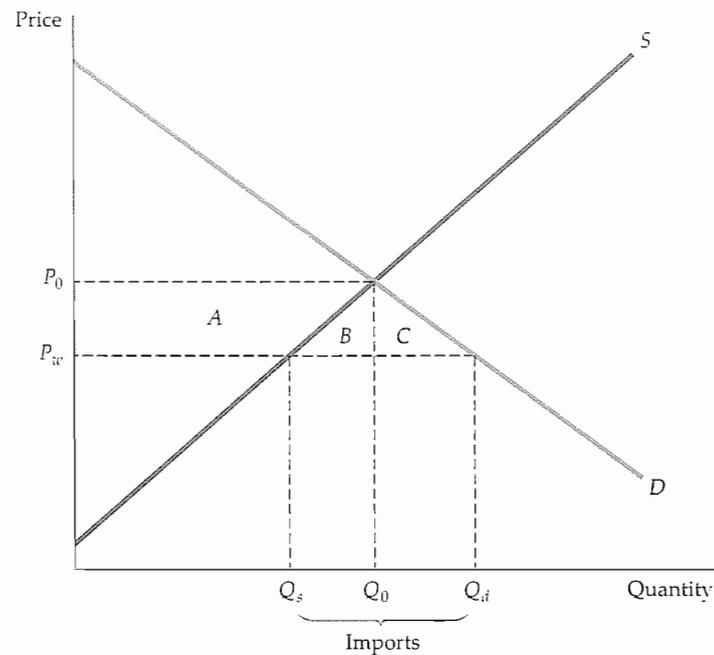


FIGURE 9.14 Import Tariff or Quota That Eliminates Imports

In a free market, the domestic price equals the world price P_w . A total Q_d is consumed, of which Q_s is supplied domestically and the rest imported. When imports are eliminated, the price is increased to P_0 . The gain to producers is trapezoid A . The loss to consumers is $A + B + C$, so the deadweight loss is $B + C$.

Imports could also be reduced to zero by imposing a sufficiently large tariff. The tariff would have to be equal to or greater than the difference between P_0 and P_w . With a tariff of this size, there will be no imports and, therefore, no government revenue from tariff collections, so the effect on consumers and producers would be the same as with a quota.

More often, government policy is designed to reduce but not eliminate imports. Again, this can be done with either a tariff or a quota, as Figure 9.15 shows. Under free trade, the domestic price will equal the world price P_w , and imports will be $Q_d - Q_s$. Now suppose a tariff of T dollars per unit is imposed on imports. Then the domestic price will rise to P^* (the world price plus the tariff); domestic production will rise and domestic consumption will fall.

In Figure 9.15, this tariff leads to a change of consumer surplus given by

$$\Delta CS = -A - B - C - D$$

The change in producer surplus is again

$$\Delta PS = A$$

Finally, the government will collect revenue in the amount of the tariff times the quantity of imports, which is rectangle D . The total change in welfare, ΔCS plus ΔPS plus the revenue to the government, is therefore $-A - B - C - D + A + D = -B - C$. Triangles B and C again represent the deadweight loss from

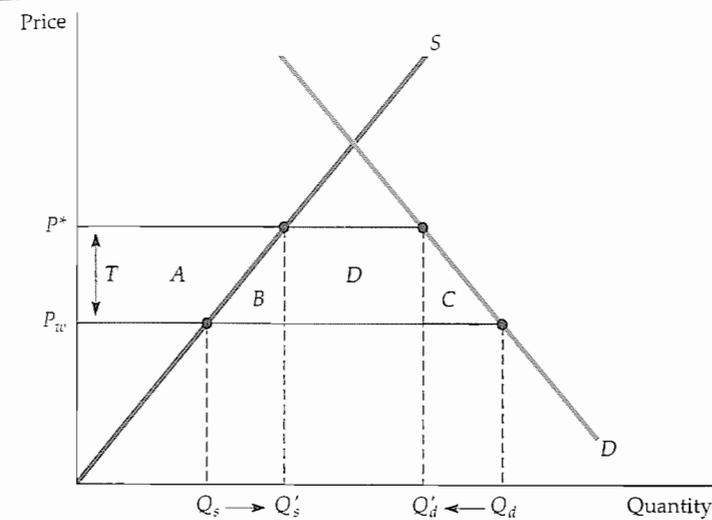


FIGURE 9.15 Import Tariff or Quota (General Case)

When imports are reduced, the domestic price is increased from P_w to P^* . This can be achieved by a quota, or by a tariff $T = P^* - P_w$. Trapezoid A is again the gain to domestic producers. The loss to consumers is $A + B + C + D$. If a tariff is used, the government gains D , the revenue from the tariff, so the net domestic loss is $B + C$. If a quota is used instead, rectangle D becomes part of the profits of foreign producers, and the net domestic loss is $B + C + D$.

restricting imports. (B represents the loss from domestic overproduction and C the loss from too little consumption.)

Suppose the government uses a quota instead of a tariff to restrict imports: Foreign producers can only ship a specific quantity ($Q'_d - Q'_s$ in Figure 9.15) to the United States and can then charge the higher price P^* for their U.S. sales. The changes in U.S. consumer and producer surplus will be the same as with the tariff, but instead of the U.S. government collecting the revenue given by rectangle D , this money will go to the foreign producers as higher profits. The United States as a whole will be even worse off than it was under the tariff, losing D as well as the deadweight loss B and C .¹⁰

This is exactly what happened with automobile imports from Japan in the 1980s. Under pressure from domestic automobile producers, the Reagan administration negotiated "voluntary" import restraints, under which the Japanese agreed to restrict shipments of cars to the United States. The Japanese could therefore sell those cars that were shipped at a price higher than the world level and capture a higher profit margin on each one. The United States would have been better off by simply imposing a tariff on these imports.

¹⁰ Alternatively, an import quota can be maintained by rationing imports to U.S. importing firms or trading companies. These middlemen would have the rights to import a fixed amount of the good each year. These rights are valuable because the middleman can buy the product on the world market at price P_w and then sell it at price P^* . The aggregate value of these rights is, therefore, given by rectangle D . If the government sells the rights for this amount of money, it can capture the same revenue it would receive with a tariff. But if these rights are given away, as sometimes happens, the money becomes a windfall to middlemen.

EXAMPLE 9.5 The Sugar Quota

In recent years the world price of sugar has been as low as 4 cents per pound, while the United States price has been 20 to 25 cents per pound. Why? By restricting imports, the U.S. government protects the \$3 billion domestic sugar industry, which would virtually be put out of business if it had to compete with low-cost foreign producers. This policy has been good for U.S. sugar producers. It has even been good for some foreign sugar producers—in particular, those whose successful lobbying efforts have given them big shares of the quota. But like most policies of this sort, it has been bad for consumers.

To see just how bad, let's look at the sugar market in 1997. Here are the relevant data for that year:

U.S. production:	15.6 billion pounds
U.S. consumption:	21.1 billion pounds
U.S. price:	21.9 cents per pound
World price:	11.1 cents per pound

At these prices and quantities, the price elasticity of U.S. supply is 1.5, and the price elasticity of U.S. demand is -0.3 .¹¹

We will fit linear supply and demand curves to these data, and then use them to calculate the effects of the quotas. You can verify that the following U.S. supply curve is consistent with a production level of 15.6 billion pounds, a price of 22 cents per pound, and a supply elasticity of 1.5:¹²

$$\text{U.S. supply: } Q_s = -7.83 + 1.07P$$

where quantity is measured in billions of pounds and price in cents per pound. Similarly, the -0.3 demand elasticity, together with the data for U.S. consumption and U.S. price, give the following linear demand curve:

$$\text{U.S. demand: } Q_D = 27.45 - 0.29P$$

These supply and demand curves are plotted in Figure 9.16. At the 11 cent world price, U.S. production would have been only about 4.0 billion pounds and U.S. consumption about 24 billion pounds, most of this imports. But fortunately for U.S. producers, imports were limited to only 5.5 billion pounds, which pushed the U.S. price up to 22 cents.

What did this cost U.S. consumers? The lost consumer surplus is given by the sum of trapezoid A, triangles B and C, and rectangle D. You should go through the calculations to verify that trapezoid A is equal to \$1078 million, triangle B to \$638 million, triangle C to \$171 million, and rectangle D to \$600 million. The total cost to consumers in 1997 was about \$2.4 billion.

How much did producers gain from this policy? Their increase in surplus is given by trapezoid A (i.e., about \$1 billion). The \$600 million of rectangle D was

¹¹ These elasticity estimates are based on Morris E. Morkre and David G. Tarr, *Effects of Restrictions on United States Imports: Five Case Studies and Theory*, U.S. Federal Trade Commission Staff Report, June 1981; and F. M. Scherer, "The United States Sugar Program," Kennedy School of Government Case Study, Harvard University, 1992. For a general discussion of sugar quotas and other aspects of U.S. agricultural policy, see D. Gale Johnson, *Agricultural Policy and Trade* (New York: New York University Press, 1985); and Gail L. Cramer and Clarence W. Jensen, *Agricultural Economics and Agribusiness* (New York: Wiley, 1985).

¹² See Section 2.6 in Chapter 2 to review the procedure for fitting linear supply and demand functions to data of this kind.

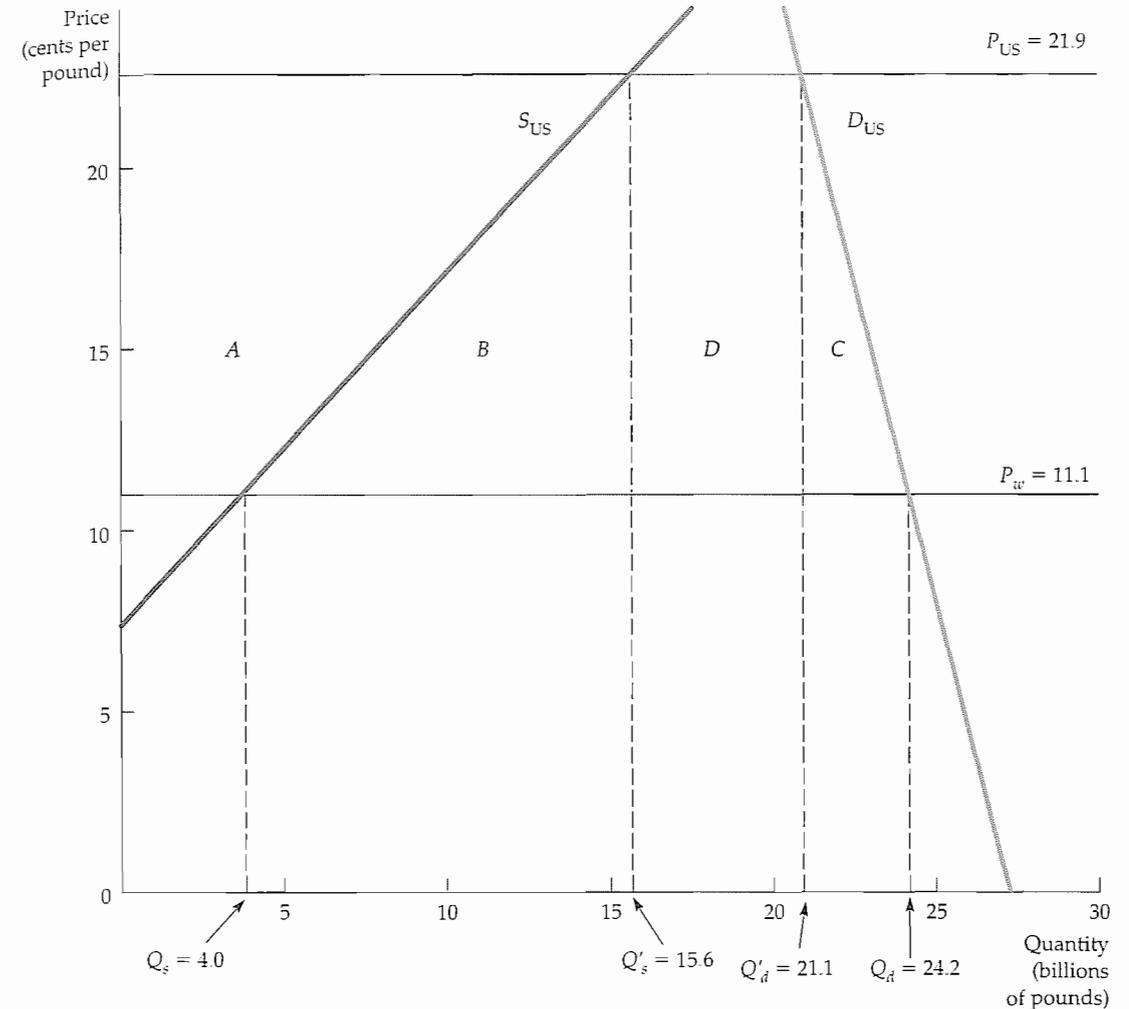


FIGURE 9.16 Sugar Quota in 1997

At the world price of 11.1 cents per pound, about 24.2 billion pounds of sugar would have been consumed in the United States in 1997, of which all but 4 billion pounds would have been imported. Restricting imports to 5.5 billion pounds caused the U.S. price to go up to 21.9 cents. The cost to consumers, $A + B + C + D$, was about \$2.4 billion. The gain to domestic producers was trapezoid A, about \$1 billion. Rectangle D, \$600 million, was a gain to those foreign producers who obtained quota allotments. Triangles B and C represent the deadweight loss of about \$800 million.

a gain for those foreign producers who succeeded in obtaining large allotments of the quota because they received a higher price for their sugar. Triangles B and C represent a deadweight loss of about \$800 million.

9.6 The Impact of a Tax or Subsidy

What would happen to the price of widgets if the government imposed a \$1 tax on every widget sold? Many people would answer that the price would increase by a dollar, with consumers now paying a dollar more per widget than they would have paid without the tax. But this answer is wrong.

Or consider the following question. The government wants to impose a 50-cent-per-gallon tax on gasoline and is considering two methods of collecting it. Under Method 1, the owner of each gas station would deposit the tax money (50 cents times the number of gallons sold) in a locked box, to be collected by a government agent. Under Method 2 the buyer would pay the tax (50 cents times the number of gallons purchased) directly to the government. Which method costs the buyer more? Many people would say Method 2, but this answer is also wrong.

The burden of a tax (or the benefit of a subsidy) falls partly on the consumer and partly on the producer. Furthermore, it does not matter who puts the money in the collection box (or sends the check to the government)—Methods 1 and 2 above both cost the consumer the same amount of money. As we will see, the share of a tax borne by consumers depends on the shapes of the supply and demand curves and, in particular, on the relative elasticities of supply and demand. As for our first question, a \$1 tax on widgets would indeed cause the price to rise, but usually by *less* than a dollar and sometimes by *much* less. To understand why, let's use supply and demand curves to see how consumers and producers are affected when a tax is imposed on a product, and what happens to price and quantity.

specific tax Tax of a certain amount of money per unit sold.

The Effects of a Specific Tax For simplicity we will consider a **specific tax**—a tax of a certain amount of money *per unit sold*. This is in contrast to an *ad valorem* (i.e., proportional) tax, such as a state sales tax. (The analysis of an ad valorem tax is roughly the same and yields the same qualitative results.) Examples of specific taxes include federal and state taxes on gasoline and cigarettes.

Suppose the government imposes a tax of t cents per unit on widgets. Assuming that everyone obeys the law, the government must then receive t cents for every widget sold. This means that the price the buyer pays must exceed the net price the seller receives by t cents. Figure 9.17 illustrates this simple accounting relationship—and its implications. Here, P_0 and Q_0 represent the market price and quantity *before* the tax is imposed. P_b is the price that buyers pay, and P_s is the net price that sellers receive *after* the tax is imposed. Note that $P_b - P_s = t$, so the government is happy.

How do we determine what the market quantity will be after the tax is imposed, and how much of the tax is borne by buyers and how much by sellers? First, remember that what buyers care about is the price that they must pay: P_b . The amount that they will buy is given by the demand curve; it is the quantity that we read off of the demand curve given a price P_b . Similarly, sellers care about the net price they receive, P_s . Given P_s , the quantity they will produce and sell is read off the supply curve. Finally, we know that the quantity that is sold must equal the quantity that is bought. The solution, then, is to find the quantity that corresponds to a price of P_b on the demand curve, and a price of P_s on the supply curve, such that the difference $P_b - P_s$ is equal to the tax t . In Figure 9.17 this quantity is shown as Q_1 .

Who bears the burden of the tax? In Figure 9.17, this burden is shared roughly equally by buyers and sellers. The market price (the price buyers pay) rises by half of the tax. And the price that sellers receive falls by roughly half of the tax.

As Figure 9.17 shows, *four conditions* must be satisfied after the tax is in place:

1. The quantity sold and the buyer's price P_b must lie on the demand curve (because buyers are interested only in the price they must pay).
2. The quantity sold and the seller's price P_s must lie on the supply curve (because sellers are concerned only with the amount of money they receive net of the tax).

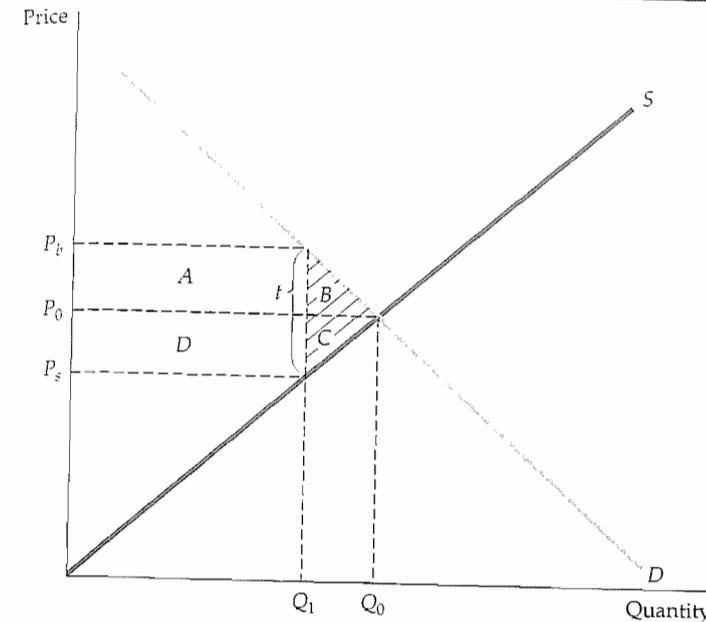


FIGURE 9.17 Incidence of a Tax

P_b is the price (including the tax) paid by buyers. P_s is the price that sellers receive, net of the tax. Here the burden of the tax is split about evenly between buyers and sellers. Buyers lose $A + B$, sellers lose $D + C$, and the government earns $A + D$ in revenue. The deadweight loss is $B + C$.

3. The quantity demanded must equal the quantity supplied (Q_1 in the figure).
4. The difference between the price the buyer pays and the price the seller receives must equal the tax t .

These conditions can be summarized by the following four equations:

$$Q^D = Q^D(P_b) \quad (9.1a)$$

$$Q^S = Q^S(P_s) \quad (9.1b)$$

$$Q^D = Q^S \quad (9.1c)$$

$$P_b - P_s = t \quad (9.1d)$$

If we know the demand curve $Q^D(P_b)$, the supply curve $Q^S(P_s)$, and the size of the tax t , we can solve these equations for the buyers' price P_b , the sellers' price P_s , and the total quantity demanded and supplied. This task is not as difficult as it may seem, as we demonstrate in Example 9.6.

Figure 9.17 also shows that a tax results in a *deadweight loss*. Because buyers pay a higher price, there is a change in consumer surplus given by

$$\Delta CS = -A - B$$

Because sellers now receive a lower price, there is also a change in producer surplus given by

$$\Delta PS = -C - D$$

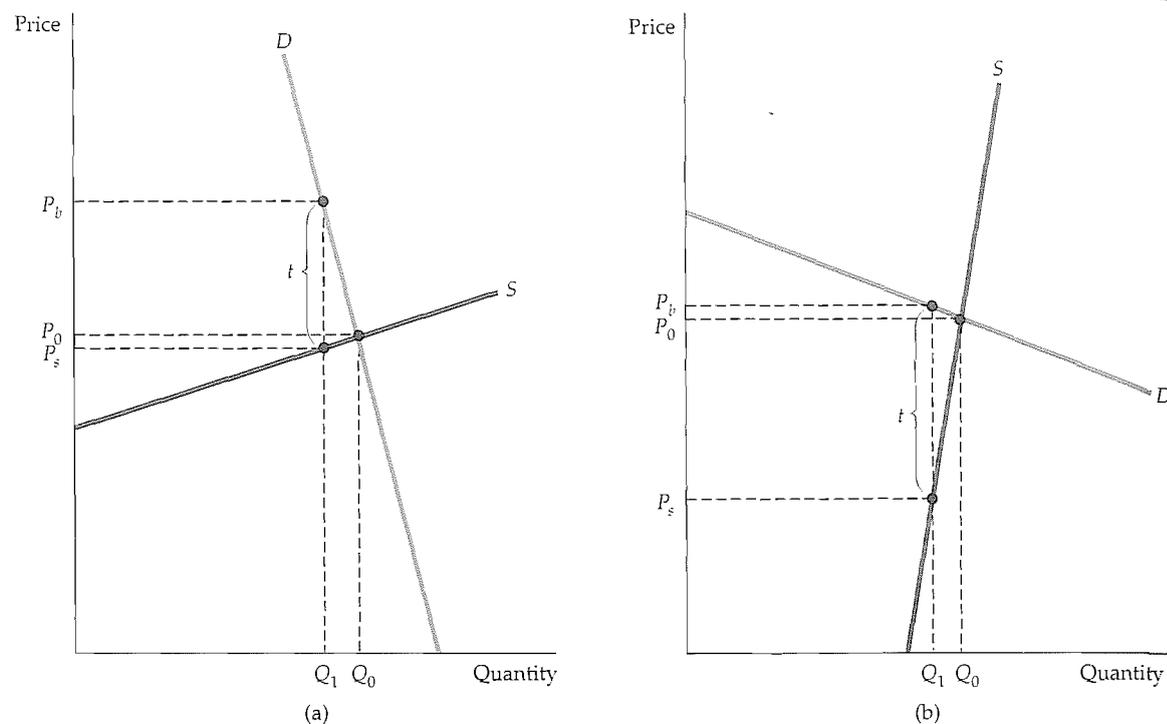


FIGURE 9.18 Impact of a Tax Depends on Elasticities of Supply and Demand

(a) If demand is very inelastic relative to supply, the burden of the tax falls mostly on buyers. (b) If demand is very elastic relative to supply, it falls mostly on sellers.

Government tax revenue is tQ_1 , the sum of rectangles A and D. The total change in welfare, ΔCS plus ΔPS plus the revenue to the government, is therefore $-A - B - C - D + A + D = -B - C$. Triangles B and C represent the dead-weight loss from the tax.

In Figure 9.17, the burden of the tax is shared almost evenly between buyers and sellers, but this is not always the case. If demand is relatively inelastic and supply is relatively elastic, the burden of the tax will fall mostly on buyers. Figure 9.18(a) shows why: It takes a relatively large increase in price to reduce the quantity demanded by even a small amount, whereas only a small price decrease is needed to reduce the quantity supplied. For example, because cigarettes are addictive, the elasticity of demand is small (about -0.3), so federal and state cigarette taxes are borne largely by cigarette buyers.¹³ Figure 9.18(b) shows the opposite case: If demand is relatively elastic and supply is relatively inelastic, the burden of the tax will fall mostly on sellers.

So even if we have only estimates of the elasticities of demand and supply at a point or for a small range of prices and quantities, instead of the entire demand and supply curves, we can still roughly determine who will bear the greatest burden of a tax (whether the tax is actually in effect or is only under discussion as a policy option). In general, a tax falls mostly on the buyer if E_d/E_s is small, and mostly on the seller if E_d/E_s is large.

¹³ See Daniel A. Sumner and Michael K. Wohlgenant, "Effects of an Increase in the Federal Excise Tax on Cigarettes," *American Journal of Agricultural Economics* 67 (May 1985): 235-42.

In fact, by using the following "pass-through" formula, we can calculate the percentage of the tax borne by buyers:

$$\text{Pass-through fraction} = E_s / (E_s - E_d)$$

This formula tells us what fraction of the tax is passed through to consumers in the form of higher prices. For example, when demand is totally inelastic, so that E_d is zero, the pass-through fraction is 1, and all the tax is borne by consumers. When demand is totally elastic, the pass-through fraction is zero, and producers bear all the tax. (The fraction of the tax producers bear is given by $-E_d / (E_s - E_d)$.)

The Effects of a Subsidy

A subsidy can be analyzed in much the same way as a tax—in fact, you can think of a subsidy as a *negative tax*. With a subsidy, the sellers' price *exceeds* the buyers' price, and the difference between the two is the amount of the subsidy. As you would expect, the effect of a subsidy on the quantity produced and consumed is just the opposite of the effect of a tax—the quantity will increase.

Figure 9.19 illustrates this. At the presubsidy market price P_0 , the elasticities of supply and demand are roughly equal. As a result, the benefit of the subsidy is shared roughly equally between buyers and sellers. As with a tax, this is not always the case. In general, the benefit of a subsidy accrues mostly to buyers if E_d/E_s is small and mostly to sellers if E_d/E_s is large.

As with a tax, given the supply curve, the demand curve, and the size of the subsidy s , we can solve for the resulting prices and quantity. The same four conditions apply for a subsidy as for a tax, but now the difference between the sellers' price and the buyers' price is equal to the subsidy. Again, we can write these conditions algebraically:

$$Q^D = Q^D(P_b) \tag{9.2a}$$

$$Q^S = Q^S(P_s) \tag{9.2b}$$

$$Q^D = Q^S \tag{9.2c}$$

$$P_s - P_b = s \tag{9.2d}$$

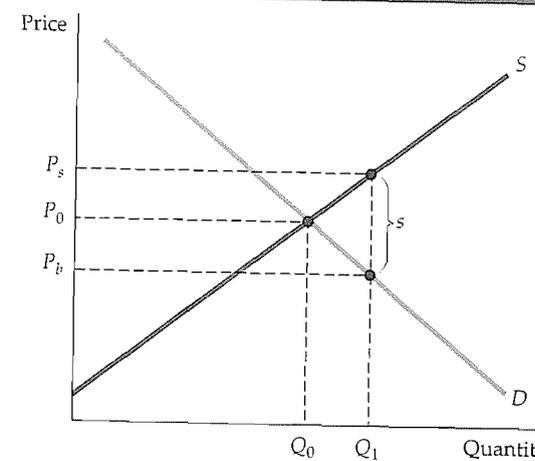


FIGURE 9.19 Subsidy

A subsidy can be thought of as a negative tax. Like a tax, the benefit of a subsidy is split between buyers and sellers, depending on the relative elasticities of supply and demand.

subsidy Payment reducing the buyer's price below the seller's price; i.e., a negative tax.

To make sure you understand how to analyze the impact of a tax or subsidy, you might find it helpful to work through one or two examples, such as Exercises 2 and 14 at the end of this chapter.

EXAMPLE 9.6 A Tax on Gasoline

The idea of a large tax on gasoline, both to raise government revenue and to reduce oil consumption and U.S. dependence on oil imports, has been discussed for many years. Let's see how a 50-cent-per-gallon tax would affect the price and consumption of gasoline.

We will do this analysis in the setting of market conditions during the mid-1990s—when gasoline was selling for about \$1 per gallon and total consumption was about 100 billion gallons per year (bg/yr).¹⁴ We will also use intermediate-run elasticities: elasticities that would apply to a period of about three to six years after a price change.

A reasonable number for the intermediate-run elasticity of gasoline demand is -0.5 (see Example 2.5 in Chapter 2). We can use this elasticity figure, together with the \$1 and 100 bg/yr price and quantity numbers, to calculate a linear demand curve for gasoline. You can verify that the following demand curve fits these data:

$$\text{Gasoline demand: } Q^D = 150 - 50P$$

Gasoline is refined from crude oil, some of which is produced domestically and some imported. (Some gasoline is also imported directly.) The supply curve for gasoline will therefore depend on the world price of oil, on domestic oil supply, and on the cost of refining. The details are beyond the scope of this example, but a reasonable number for the elasticity of supply is 0.4 . You should verify that this elasticity, together with the \$1 and 100 bg/yr price and quantity, gives the following linear supply curve:

$$\text{Gasoline supply: } Q^S = 60 + 40P$$

You should also verify that these demand and supply curves imply a market price of \$1 and quantity of 100 bg/yr.

We can use these linear demand and supply curves to calculate the effect of a 50-cent-per-gallon tax. First, we write the four conditions that must hold, as given by equations (9.1a–d):

$$\begin{aligned} Q^D &= 150 - 50P_b && \text{(Demand)} \\ Q^S &= 60 + 40P_s && \text{(Supply)} \\ Q^D &= Q^S && \text{(Supply must equal demand)} \\ P_b - P_s &= 0.50 && \text{(Government must receive 50 cents/gallon)} \end{aligned}$$

Now combine the first three equations to equate supply and demand:

$$150 - 50P_b = 60 + 40P_s$$

In §2.5, we explain that demand is often more price elastic in the long run than in the short run because it takes time for people to change their consumption habits and/or because the demand for a good might be linked to the stock of another good that changes slowly.

For a review of the procedure for calculating linear curves, see §2.5. Given data for price and quantity, as well as estimates of demand and supply elasticities, we can use a two-step procedure to solve for quantity demanded and supplied.

¹⁴Of course, this price varied across regions and grades of gasoline, but we can ignore this here. Quantities of oil and oil products are often measured in barrels; there are 42 gallons in a barrel, so the quantity figure could also be written as 2.4 billion barrels per year.

We can rewrite the last of the four equations as $P_b = P_s + 0.50$ and substitute this for P_b in the above equation:

$$150 - 50(P_s + 0.50) = 60 + 40P_s$$

Now we can rearrange this equation and solve for P_s :

$$50P_s + 40P_s = 150 - 25 - 60$$

$$90P_s = 65, \text{ or } P_s = .72$$

Remember that $P_b = P_s + 0.50$, so $P_b = 1.22$. Finally, we can determine the total quantity from either the demand or supply curve. Using the demand curve (and the price $P_b = 1.22$), we find that $Q = 150 - (50)(1.22) = 150 - 61$, or $Q = 89$ bg/yr. This represents an 11-percent decline in gasoline consumption. Figure 9.20 illustrates these calculations and the effect of the tax.

The burden of this tax would be split roughly evenly between consumers and producers. Consumers would pay about 22 cents per gallon more for gasoline, and producers would receive about 28 cents per gallon less. It should not be surprising then, that both consumers and producers opposed such a tax, and politicians representing both groups fought the proposal every time it came up. But note that the tax would raise significant revenue for the government. The annual revenue would be $tQ = (0.50)(89) = \$44.5$ billion per year.

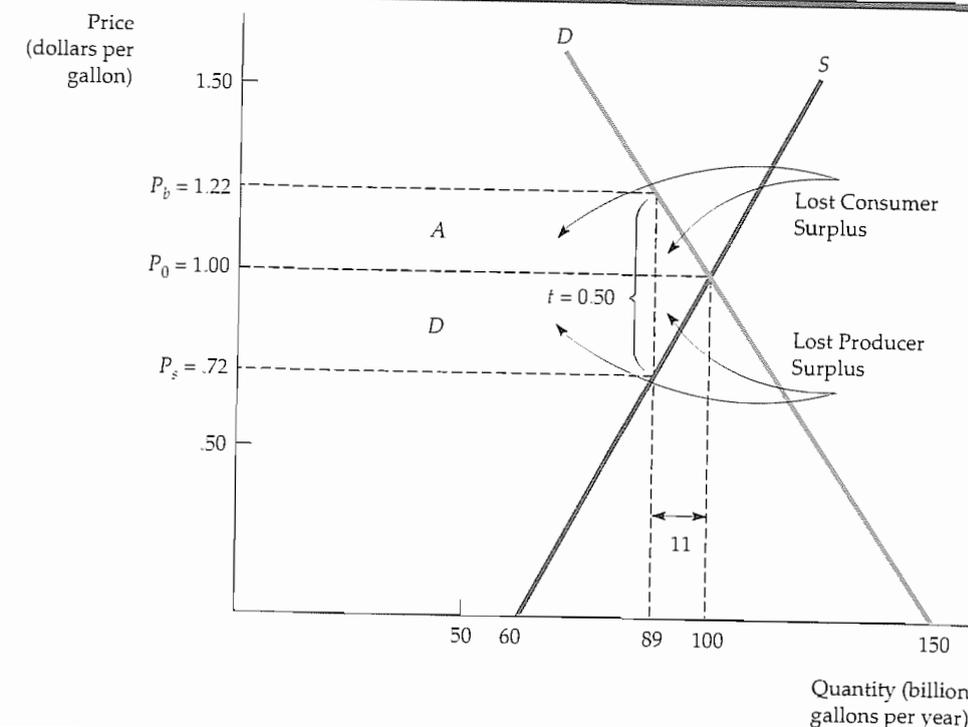


FIGURE 9.20 Impact of 50-Cent Gasoline Tax

The price of gasoline at the pump increases from \$1.00 per gallon to \$1.22, and the quantity sold falls from 100 to 89 bg/yr. Annual revenue from the tax is $(0.50)(89) = \$44.5$ billion. The two triangles show the deadweight loss of \$2.75 billion per year.

The cost to consumers and producers, however, will be more than the \$44.5 billion in tax revenue. Figure 9.20 shows the deadweight loss from this tax as the two shaded triangles. The two rectangles *A* and *D* represent the total tax collected by the government, but the total loss of consumer and producer surplus is larger.

Before deciding whether a gasoline tax is desirable, it is important to know how large the resulting deadweight loss is likely to be. We can easily calculate this from Figure 9.20. Combining the two small triangles into one large one, we see that the area is

$$\begin{aligned} & (1/2) \times (\$0.50/\text{gallon}) \times (11 \text{ billion gallons/year}) \\ & = \$2.75 \text{ billion per year} \end{aligned}$$

This deadweight loss is about 6 percent of the government revenue resulting from the tax, and must be balanced against any additional benefits that the tax might bring.

SUMMARY

- Simple models of supply and demand can be used to analyze a wide variety of government policies. Specific policies that we have examined include price controls, minimum prices, price support programs, production quotas or incentive programs to limit output, import tariffs and quotas, and taxes and subsidies.
- In each case, consumer and producer surplus are used to evaluate the gains and losses to consumers and producers. Applying the methodology to natural gas price controls, airline regulation, price supports for wheat, and the sugar quota, we found that these gains and losses can be quite large.
- When government imposes a tax or subsidy, price usually does not rise or fall by the full amount of the tax or subsidy. Also, the incidence of a tax or subsidy is usually split between producers and consumers. The fraction that each group ends up paying or receiving depends on the relative elasticities of supply and demand.
- Government intervention generally leads to a deadweight loss; even if consumer surplus and producer surplus are weighted equally, there will be a net loss from government policies that shifts surplus from one group to the other. In some cases this deadweight loss will be small, but in other cases—price supports and import quotas are examples—it is large. This deadweight loss is a form of economic inefficiency that must be taken into account when policies are designed and implemented.
- Government intervention in a competitive market is not always bad. Government—and the society it represents—might have objectives other than economic efficiency. And there are situations in which government intervention can improve economic efficiency. Examples are externalities and cases of market failure. These situations, and the way government can respond to them, are discussed in Chapters 17 and 18.

QUESTIONS FOR REVIEW

- What is meant by *deadweight loss*? Why does a price ceiling usually result in a deadweight loss?
- Suppose the supply curve for a good is completely inelastic. If the government imposed a price ceiling below the market-clearing level, would a deadweight loss result? Explain.
- How can a price ceiling make consumers better off? Under what conditions might it make them worse off?
- Suppose the government regulates the price of a good to be no lower than some minimum level. Can such a minimum price make producers as a whole worse off? Explain.
- How are production limits used in practice to raise the prices of the following goods or services: (a) taxi rides, (b) drinks in a restaurant or bar, (c) wheat or corn?

- Suppose the government wants to increase farmers' incomes. Why do price supports or acreage limitation programs cost society more than simply giving farmers money?
- Suppose the government wants to limit imports of a certain good. Is it preferable to use an import quota or a tariff? Why?

- The burden of a tax is shared by producers and consumers. Under what conditions will consumers pay most of the tax? Under what conditions will producers pay most of it? What determines the share of a subsidy that benefits consumers?
- Why does a tax create a deadweight loss? What determines the size of this loss?

EXERCISES

- In 1996, the U.S. Congress raised the minimum wage from \$4.25 per hour to \$5.15 per hour. Some people suggested that a government subsidy could help employers finance the higher wage. This exercise examines the economics of a minimum wage and wage subsidies. Suppose the supply of low-skilled labor is given by

$$L^S = 10w$$

where L^S is the quantity of low-skilled labor (in millions of persons employed each year), and w is the wage rate (in dollars per hour). The demand for labor is given by

$$L^D = 80 - 10w$$

- What will the free-market wage rate and employment level be? Suppose the government sets a minimum wage of \$5 per hour. How many people would then be employed?
 - Suppose that instead of a minimum wage, the government pays a subsidy of \$1 per hour for each employee. What will the total level of employment be now? What will the equilibrium wage rate be?
- Suppose the market for widgets can be described by the following equations:

$$\text{Demand: } P = 10 - Q$$

$$\text{Supply: } P = Q - 4$$

where P is the price in dollars per unit and Q is the quantity in thousands of units. Then

- What is the equilibrium price and quantity?
- Suppose the government imposes a tax of \$1 per unit to reduce widget consumption and raise government revenues. What will the new equilibrium quantity be? What price will the buyer pay? What amount per unit will the seller receive?
- Suppose the government has a change of heart about the importance of widgets to the happiness of the American public. The tax is removed and a subsidy of \$1 per unit granted to widget producers. What will the equilibrium quantity be? What

price will the buyer pay? What amount per unit (including the subsidy) will the seller receive? What will be the total cost to the government?

- Japanese rice producers have extremely high production costs, in part due to the high opportunity cost of land and to their inability to take advantage of economies of large-scale production. Analyze two policies intended to maintain Japanese rice production: (1) a per-pound subsidy to farmers for each pound of rice produced, or (2) a per-pound tariff on imported rice. Illustrate with supply-and-demand diagrams the equilibrium price and quantity, domestic rice production, government revenue or deficit, and deadweight loss from each policy. Which policy is the Japanese government likely to prefer? Which policy are Japanese farmers likely to prefer?
- In 1983, the Reagan administration introduced a new agricultural program called the Payment-in-Kind Program. To see how the program worked, let's consider the wheat market.
 - Suppose the demand function is $Q^D = 28 - 2P$ and the supply function is $Q^S = 4 + 4P$, where P is the price of wheat in dollars per bushel, and Q is the quantity in billions of bushels. Find the free-market equilibrium price and quantity.
 - Now suppose the government wants to lower the supply of wheat by 25 percent from the free-market equilibrium by paying farmers to withdraw land from production. However, the payment is made in wheat rather than in dollars—hence the name of the program. The wheat comes from the government's vast reserves that resulted from previous price support programs. The amount of wheat paid is equal to the amount that could have been harvested on the land withdrawn from production. Farmers are free to sell this wheat on the market. How much is now produced by farmers? How much is indirectly supplied to the market by the government? What is the new market price? How much do farmers gain? Do consumers gain or lose?
 - Had the government not given the wheat back to the farmers, it would have stored or destroyed it. Do taxpayers gain from the program? What potential problems does the program create?

5. About 100 million pounds of jelly beans are consumed in the United States each year, and the price has been about 50 cents per pound. However, jelly bean producers feel that their incomes are too low and have convinced the government that price supports are in order. The government will therefore buy up as many jelly beans as necessary to keep the price at \$1 per pound. However, government economists are worried about the impact of this program because they have no estimates of the elasticities of jelly bean demand or supply.
 - a. Could this program cost the government *more* than \$50 million per year? Under what conditions? Could it cost *less* than \$50 million per year? Under what conditions? Illustrate with a diagram.
 - b. Could this program cost consumers (in terms of lost consumer surplus) *more* than \$50 million per year? Under what conditions? Could it cost consumers *less* than \$50 million per year? Under what conditions? Again, use a diagram to illustrate.
6. In Exercise 3 of Chapter 2, we examined a vegetable fiber traded in a competitive world market and imported into the United States at a world price of \$9 per pound. U.S. domestic supply and demand for various price levels are shown in the following table.

PRICE	U.S. SUPPLY (MILLION POUNDS)	U.S. DEMAND (MILLION POUNDS)
3	2	34
6	4	28
9	6	22
12	8	16
15	10	10
18	12	4

Answer the following about the U.S. market:

- a. Confirm that the demand curve is given by $Q_D = 40 - 2P$, and that the supply curve is given by $Q_S = 2/3P$.
 - b. Confirm that if there were no restrictions on trade, the United States would import 16 million pounds.
 - c. If the United States imposes a tariff of \$9 per pound, what will be the U.S. price and level of imports? How much revenue will the government earn from the tariff? How large is the deadweight loss?
 - d. If the United States has no tariff but imposes an import quota of 8 million pounds, what will be the U.S. domestic price? What is the cost of this quota for U.S. consumers of the fiber? What is the gain for U.S. producers?
7. A particular metal is traded in a highly competitive world market at a world price of \$9 per ounce. Unlimited quantities are available for import into the

United States at this price. The supply of this metal from domestic U.S. mines and mills can be represented by the equation $Q^S = 2/3P$, where Q^S is U.S. output in million ounces and P is the domestic price. The demand for the metal in the United States is $Q^D = 40 - 2P$, where Q^D is the domestic demand in million ounces.

In recent years the U.S. industry has been protected by a tariff of \$9 per ounce. Under pressure from other foreign governments, the United States plans to reduce this tariff to zero. Threatened by this change, the U.S. industry is seeking a Voluntary Restraint Agreement that would limit imports into the United States to 8 million ounces per year.

- a. Under the \$9 tariff, what was the U.S. domestic price of the metal?
 - b. If the United States eliminates the tariff and the Voluntary Restraint Agreement is approved, what will be the U.S. domestic price of the metal?
8. Among the tax proposals regularly considered by Congress is an additional tax on distilled liquors. The tax would not apply to beer. The price elasticity of supply of liquor is 4.0, and the price elasticity of demand is -0.2 . The cross-elasticity of demand for beer with respect to the price of liquor is 0.1.
- a. If the new tax is imposed, who will bear the greater burden—liquor suppliers or liquor consumers? Why?
 - b. Assuming that beer supply is infinitely elastic, how will the new tax affect the beer market?
9. In Example 9.1, we calculated the gains and losses from price controls on natural gas and found that there was a deadweight loss of \$1.4 billion. This calculation was based on a price of oil of \$8 per barrel. If the price of oil were \$12 per barrel, what would the free-market price of gas be? How large a deadweight loss would result if the maximum allowable price of natural gas were \$1.00 per thousand cubic feet?
10. Example 9.5 describes the effects of the sugar quota. In 1997, imports were limited to 5.5 billion pounds, which pushed the domestic price to 22 cents per pound. Suppose imports were expanded to 6.5 billion pounds.
- a. What would be the new U.S. domestic price?
 - b. How much would consumers gain and domestic producers lose?
 - c. What would be the effect on deadweight loss and foreign producers?
11. Review Example 9.5 on the sugar quota. During the mid-1990s, U.S. sugar producers became more efficient, causing the domestic supply curve to shift to the right. We will examine the implications of this shift. Suppose that the supply curve shifts to the right by 5.5 billion pounds, so that the new supply curve is given by

$$Q_S = -2.33 + 1.07P$$

- a. Show that if the demand curve remains the same as in Example 9.5, domestic demand will equal domestic supply at a price of 21.9 cents per pound. Thus the U.S. price could be maintained at 21.9 cents with no imports.
 - b. Suppose that under pressure from foreign sugar producers, the U.S. government allows imports of 2.5 billion pounds and requires domestic producers to reduce production by the same amount. Draw the supply and demand curves and calculate the resulting cost to consumers, the benefit to foreign and domestic producers, and the deadweight loss.
12. The domestic supply and demand curves for hula beans are as follows:
- $$\text{Supply: } P = 50 + Q$$
- $$\text{Demand: } P = 200 - 2Q$$
- where P is the price in cents per pound and Q is the quantity in millions of pounds. The U.S. is a small producer in the world hula bean market, where the current price (which will not be affected by anything we do), is 60 cents per pound. Congress is considering a tariff of 40 cents per pound. Find the domestic price of hula beans that will result if the tariff is imposed. Also compute the dollar gain or loss to domestic consumers, domestic producers, and government revenue from the tariff.
13. Currently, the social security payroll tax in the United States is evenly divided between employers and employees. Employers must pay the government a

- tax of 6.2 percent of the wages they pay, and employees must pay 6.2 percent of the wages they receive. Suppose the tax were changed so that the employers paid the full 12.4 percent, and the employees paid nothing. Would employees then be better off?
14. You know that if a tax is imposed on a particular product, the burden of the tax is shared by producers and consumers. You also know that the demand for automobiles is characterized by a stock adjustment process. Suppose a special 20-percent sales tax is suddenly imposed on automobiles. Will the share of the tax paid by consumers rise, fall, or stay the same over time? Explain briefly. Repeat for a 50-cents-per-gallon gasoline tax.
15. In 1998, Americans smoked 23.5 billion packs of cigarettes. They paid an average retail price of \$2 per pack.
- a. Given that the elasticity of supply is 0.5 and the elasticity of demand is -0.4 , derive linear demand and supply curves for cigarettes.
 - b. In November 1998, after settling a lawsuit filed by 46 states, the three major tobacco companies raised the retail price of a pack of cigarettes by 45 cents. What is the new equilibrium price and quantity? How many fewer packs of cigarettes are sold?
 - c. Cigarettes are subject to a federal tax, which was about 25 cents per pack in 1998. This tax will increase by 15 cents in 2002. What will this increase do to the market-clearing price and quantity?
 - d. How much of the federal tax will consumers pay? What part will producers pay?

PART 3

CHAPTERS

- 10 Market Power: Monopoly and Monopsony 327
- 11 Pricing with Market Power 369
- 12 Monopolistic Competition and Oligopoly 423
- 13 Game Theory and Competitive Strategy 461
- 14 Markets for Factor Inputs 501
- 15 Investment, Time, and Capital Markets 533

Market Structure and Competitive Strategy

PART 3 examines a broad range of markets and explains how the pricing, investment, and output decisions of firms depend on market structure and the behavior of competitors.

Chapters 10 and 11 examine *market power*: the ability to affect price, either by a seller or a buyer. We will see how market power arises, how it differs across firms, how it affects the welfare of consumers and producers, and how it can be limited by government. We will also see how firms can design pricing and advertising strategies to take maximum advantage of their market power.

Chapters 12 and 13 deal with markets in which the number of firms is limited. We will examine a variety of such markets, ranging from *monopolistic competition*, in which many firms sell differentiated products, to *cartels*, in which a group of firms coordinate decisions and act as a monopolist. We are particularly concerned with markets in which there are only a few firms. In these cases, each firm must design its pricing, output, and investment strategies while keeping in mind how competitors are likely to react. We will develop and apply principles from game theory to analyze such strategies.

Chapter 14 shows how markets for factor inputs, such as labor and raw materials, operate. We will examine the firm's input decisions and show how those decisions depend on the structure of the input market. Chapter 15 then focuses on capital investment decisions. We will see how a firm can value the profits it expects an investment to yield in the future, and then compare this value with the cost of the investment to determine whether the investment is worthwhile.

PART 3

CHAPTERS

- 10 Market Power: Monopoly and Monopsony 327
- 11 Pricing with Market Power 369
- 12 Monopolistic Competition and Oligopoly 423
- 13 Game Theory and Competitive Strategy 461
- 14 Markets for Factor Inputs 501
- 15 Investment, Time, and Capital Markets 533

Market Structure and Competitive Strategy

PART 3 examines a broad range of markets and explains how the pricing, investment, and output decisions of firms depend on market structure and the behavior of competitors.

Chapters 10 and 11 examine *market power*: the ability to affect price, either by a seller or a buyer. We will see how market power arises, how it differs across firms, how it affects the welfare of consumers and producers, and how it can be limited by government. We will also see how firms can design pricing and advertising strategies to take maximum advantage of their market power.

Chapters 12 and 13 deal with markets in which the number of firms is limited. We will examine a variety of such markets, ranging from *monopolistic competition*, in which many firms sell differentiated products, to *cartels*, in which a group of firms coordinate decisions and act as a monopolist. We are particularly concerned with markets in which there are only a few firms. In these cases, each firm must design its pricing, output, and investment strategies while keeping in mind how competitors are likely to react. We will develop and apply principles from game theory to analyze such strategies.

Chapter 14 shows how markets for factor inputs, such as labor and raw materials, operate. We will examine the firm's input decisions and show how those decisions depend on the structure of the input market. Chapter 15 then focuses on capital investment decisions. We will see how a firm can value the profits it expects an investment to yield in the future, and then compare this value with the cost of the investment to determine whether the investment is worthwhile.

CHAPTER 10

Market Power: Monopoly and Monopsony

In a perfectly competitive market, the large number of sellers and buyers of a good ensures that no single seller or buyer can affect its price. The market forces of supply and demand determine price. Individual firms take the market price as a given in deciding how much to produce and sell, and consumers take it as a given in deciding how much to buy.

Monopoly and *monopsony*, the subjects of this chapter, are the polar opposites of perfect competition. A **monopoly** is a market that has only one seller but many buyers. A **monopsony** is just the opposite: a market with many sellers but only one buyer. Monopoly and monopsony are closely related, which is why we cover them in the same chapter.

First we discuss the behavior of a monopolist. Because a monopolist is the sole producer of a product, the demand curve that it faces is the market demand curve. This market demand curve relates the price that the monopolist receives to the quantity it offers for sale. We will see how a monopolist can take advantage of its control over price and how the profit-maximizing price and quantity differ from what would prevail in a competitive market.

In general, the monopolist's quantity will be lower and its price higher than the competitive quantity and price. This imposes a cost on society because fewer consumers buy the product, and those who do pay more for it. This is why antitrust laws exist which forbid firms from monopolizing most markets. When economies of scale make monopoly desirable—for example, with local electric power companies—we will see how the government can then increase efficiency by regulating the monopolist's price.

Pure monopoly is rare, but in many markets only a few firms compete with each other. The interactions of firms in such markets can be complicated and often involve aspects of strategic gaming, a topic covered in Chapters 12 and 13. In any case, the firms may be able to affect price and may find it profitable to charge a price higher than marginal cost. These firms have *monopoly power*. We will discuss the determinants of monopoly power, its measurement, and its implications for pricing.

Chapter Outline

- 10.1 Monopoly 328
- 10.2 Monopoly Power 339
- 10.3 Sources of Monopoly Power 345
- 10.4 The Social Costs of Monopoly Power 347
- 10.5 Monopsony 352
- 10.6 Monopsony Power 355
- 10.7 Limiting Market Power: The Antitrust Laws 359

List of Examples

- 10.1 Astra-Merck Prices
Prilosec 334
- 10.2 Markup Pricing:
Supermarkets to Designer
Jeans 342
- 10.3 The Pricing of
Prerecorded
Videocassettes 343
- 10.4 Monopsony Power in U.S.
Manufacturing 358
- 10.5 A Phone Call About
Prices 362
- 10.6 The United States versus
Microsoft 363

monopoly Market with only one seller.

monopsony Market with only one buyer.

market power Ability of a seller or buyer to affect the price of a good.

Next we will turn to *monopsony*. Unlike a competitive buyer, a monopsonist pays a price that depends on the quantity that it purchases. The monopsonist's problem is to choose the quantity that maximizes its net benefit from the purchase—the value derived from the good less the money paid for it. By showing how the choice is made, we will demonstrate the close parallel between monopsony and monopoly.

Although pure monopsony is also unusual, many markets have only a few buyers who can purchase the good for less than they would pay in a competitive market. These buyers have *monopsony power*. Typically, this situation occurs in markets for inputs to production. For example, General Motors, the largest U.S. car manufacturer, has monopsony power in the markets for tires, car batteries, and other parts. We will discuss the determinants of monopsony power, its measurement, and its implications for pricing.

Monopoly and monopsony power are two forms of **market power**: the ability—of either a seller or a buyer—to affect the price of a good.¹ Because sellers or buyers have at least some market power (in most real-world markets), we need to understand how market power works and how it affects producers and consumers.

10.1 Monopoly

As the sole producer of a product, a monopolist is in a unique position. If the monopolist decides to raise the price of the product, it need not worry about competitors who, by charging lower prices, would capture a larger share of the market at the monopolist's expense. The monopolist *is* the market and completely controls the amount of output offered for sale.

But this does not mean that the monopolist can charge any price it wants—at least not if its objective is to maximize profit. This textbook is a case in point. Prentice Hall, Inc., owns the copyright and is, therefore, a monopoly producer of this book. Then why doesn't it sell the book for \$500 a copy? Because few people would buy it, and Prentice Hall would earn a much lower profit.

To maximize profit, the monopolist must first determine its costs and the characteristics of market demand. Knowledge of demand and cost is crucial for a firm's economic decision making. Given this knowledge, the monopolist must then decide how much to produce and sell. The price per unit that the monopolist receives then follows directly from the market demand curve. Equivalently, the monopolist can determine price, and the quantity it will sell at that price follows from the market demand curve.

Average Revenue and Marginal Revenue

The monopolist's average revenue—the price it receives per unit sold—is precisely the market demand curve. To choose its profit-maximizing output level, the monopolist also needs to know its **marginal revenue**: the change in revenue

marginal revenue Change in revenue resulting from a one-unit increase in output.

¹ The courts use the term "monopoly power" to mean significant and sustainable market power sufficient to warrant particular scrutiny under the antitrust laws. In this book, however, for pedagogic reasons we use "monopoly power" differently, to mean market power on the part of sellers whether substantial or not.

TABLE 10.1 Total, Marginal, and Average Revenue

PRICE (P)	QUANTITY (Q)	TOTAL REVENUE (R)	MARGINAL REVENUE (MR)	AVERAGE REVENUE (AR)
\$6	0	\$0	—	—
5	1	5	\$5	\$5
4	2	8	3	4
3	3	9	1	3
2	4	8	-1	2
1	5	5	-3	1

that results from a unit change in output. To see the relationship among total, average, and marginal revenue, consider a firm facing the following demand curve:

$$P = 6 - Q$$

Table 10.1 shows the behavior of total, average, and marginal revenue for this demand curve. Note that revenue is zero when the price is \$6: At that price, nothing is sold. At a price of \$5, however, one unit is sold, so total (and marginal) revenue is \$5. An increase in quantity sold from 1 to 2 increases revenue from \$5 to \$8; marginal revenue is thus \$3. As quantity sold increases from 2 to 3, marginal revenue falls to \$1, and when it increases from 3 to 4, marginal revenue becomes negative. When marginal revenue is positive, revenue is increasing with quantity, but when marginal revenue is negative, revenue is decreasing.

When the demand curve is downward sloping, the price (average revenue) is greater than marginal revenue because all units are sold at the same price. If sales are to increase by 1 unit, the price must fall. In that case, all units sold, not just the additional unit, will earn less revenue. Note, for example, what happens in Table 10.1 when output is increased from 1 to 2 units and price is reduced to \$4. Marginal revenue is \$3: \$4 (the revenue from the sale of the additional unit of output) less \$1 (the loss of revenue from selling the first unit for \$4 instead of \$5). Thus, marginal revenue (\$3) is less than price (\$4).

Figure 10.1 plots average and marginal revenue for the data in Table 10.1. Our demand curve is a straight line, and in this case, the marginal revenue curve has twice the slope of the demand curve (and the same intercept).²

The Monopolist's Output Decision

What quantity should the monopolist produce? In Chapter 8, we saw that to maximize profit, a firm must set output so that marginal revenue is equal to marginal cost. This is the solution to the monopolist's problem. In Figure 10.2, the market demand curve *D* is the monopolist's average revenue curve. It specifies the price per unit that the monopolist receives as a function of its output level. Also shown are the corresponding marginal revenue curve *MR* and the

² If the demand curve is written so that price is a function of quantity, $P = a - bQ$, total revenue is given by $PQ = aQ - bQ^2$. Marginal revenue (using calculus) is $d(PQ)/dQ = a - 2bQ$. In this example, demand is $P = 6 - Q$ and marginal revenue is $MR = 6 - 2Q$. (This holds only for small changes in *Q* and therefore does not exactly match the data in Table 10.1.)

In §8.2, we explain that marginal revenue is a measure of how much revenue increases when output increases by one unit.

In §7.2, we explain that marginal cost is the change in variable cost associated with a one-unit increase in output.

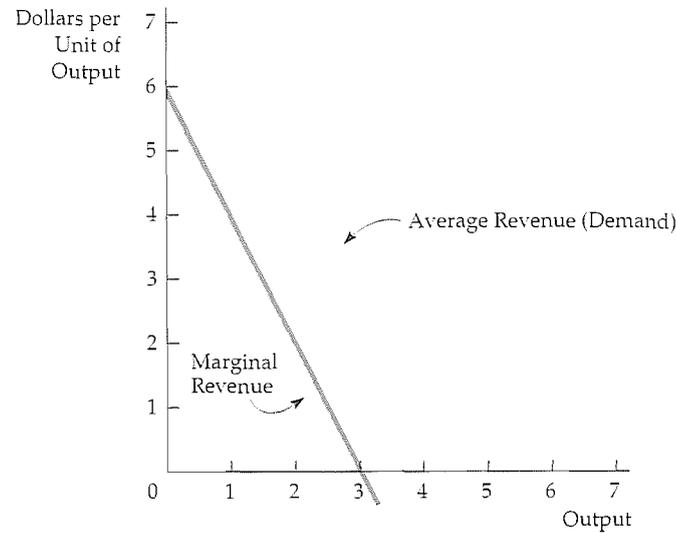


FIGURE 10.1 Average and Marginal Revenue
Average and marginal revenue are shown for the demand curve $P = 6 - Q$.

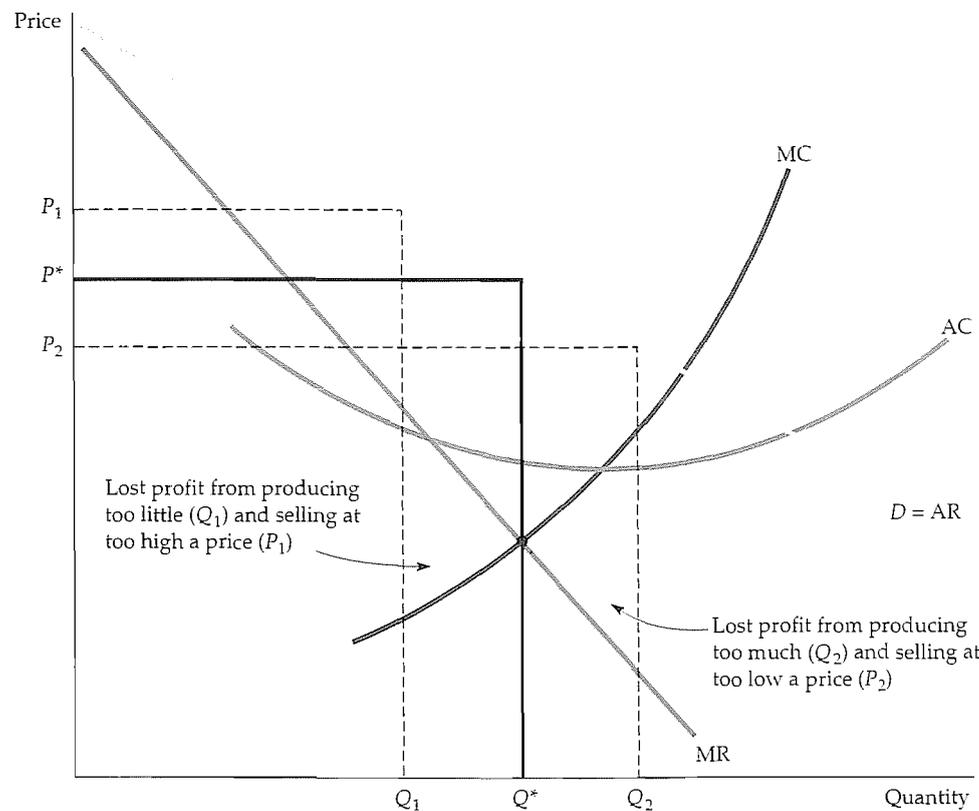


FIGURE 10.2 Profit Is Maximized When Marginal Revenue Equals Marginal Cost

Q^* is the output level at which $MR = MC$. If the firm produces a smaller output—say, Q_1 —it sacrifices some profit because the extra revenue that could be earned from producing and selling the units between Q_1 and Q^* exceeds the cost of producing them. Similarly, expanding output from Q^* to Q_2 would reduce profit because the additional cost would exceed the additional revenue.

average and marginal cost curves, AC and MC . Marginal revenue and marginal cost are equal at quantity Q^* . Then from the demand curve, we find the price P^* that corresponds to this quantity Q^* .

How can we be sure that Q^* is the profit-maximizing quantity? Suppose the monopolist produces a smaller quantity Q_1 and receives the corresponding higher price P_1 . As Figure 10.2 shows, marginal revenue would then exceed marginal cost. In that case, if the monopolist produced a little more than Q_1 , it would receive extra profit ($MR - MC$) and thereby increase its total profit. In fact, the monopolist could keep increasing output, adding more to its total profit until output Q^* , at which point the incremental profit earned from producing one more unit is zero. So the smaller quantity Q_1 is not profit maximizing, even though it allows the monopolist to charge a higher price. If the monopolist produced Q_1 instead of Q^* , its total profit would be smaller by an amount equal to the shaded area below the MR curve and above the MC curve, between Q_1 and Q^* .

In Figure 10.2, the larger quantity Q_2 is likewise not profit maximizing. At this quantity, marginal cost exceeds marginal revenue. Therefore, if the monopolist produced a little less than Q_2 , it would increase its total profit (by $MC - MR$). It could increase its profit even more by reducing output all the way to Q^* . The increased profit achieved by producing Q^* instead of Q_2 is given by the area below the MC curve and above the MR curve, between Q^* and Q_2 .

We can also see algebraically that Q^* maximizes profit. Profit π is the difference between revenue and cost, both of which depend on Q :

$$\pi(Q) = R(Q) - C(Q)$$

As Q is increased from zero, profit will increase until it reaches a maximum and then begin to decrease. Thus the profit-maximizing Q is such that the incremental profit resulting from a small increase in Q is just zero (i.e., $\Delta\pi/\Delta Q = 0$). Then

$$\Delta\pi/\Delta Q = \Delta R/\Delta Q - \Delta C/\Delta Q = 0$$

But $\Delta R/\Delta Q$ is marginal revenue and $\Delta C/\Delta Q$ is marginal cost. Thus the profit-maximizing condition is that $MR - MC = 0$, or $MR = MC$.

An Example

To grasp this result more clearly, let's look at an example. Suppose the cost of production is

$$C(Q) = 50 + Q^2$$

In other words, there is a fixed cost of \$50, and variable cost is Q^2 . Suppose demand is given by

$$P(Q) = 40 - Q$$

By setting marginal revenue equal to marginal cost, you can verify that profit is maximized when $Q = 10$, an output level that corresponds to a price of \$30.³

³ Note that average cost is $C(Q)/Q = 50/Q + Q$ and marginal cost is $\Delta C/\Delta Q = 2Q$. Revenue is $R(Q) = P(Q)Q = 40Q - Q^2$, so marginal revenue is $MR = \Delta R/\Delta Q = 40 - 2Q$. Setting marginal revenue equal to marginal cost gives $40 - 2Q = 2Q$, or $Q = 10$.

Cost, revenue, and profit are plotted in Figure 10.3(a). When the firm produces little or no output, profit is negative because of the fixed cost. Profit increases as Q increases, reaching a maximum of \$150 at $Q^* = 10$, and then decreases as Q is increased further. And at the point of maximum profit, the slopes of the revenue and cost curves are the same. (Note that the tangent lines rr' and cc' are parallel.)

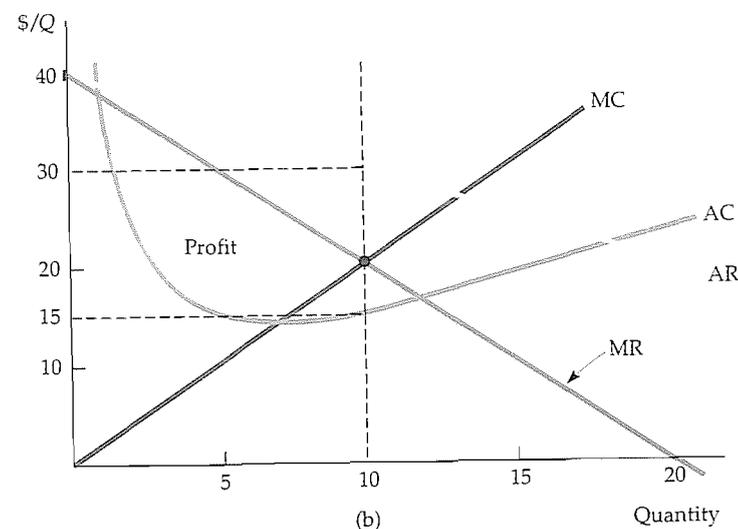
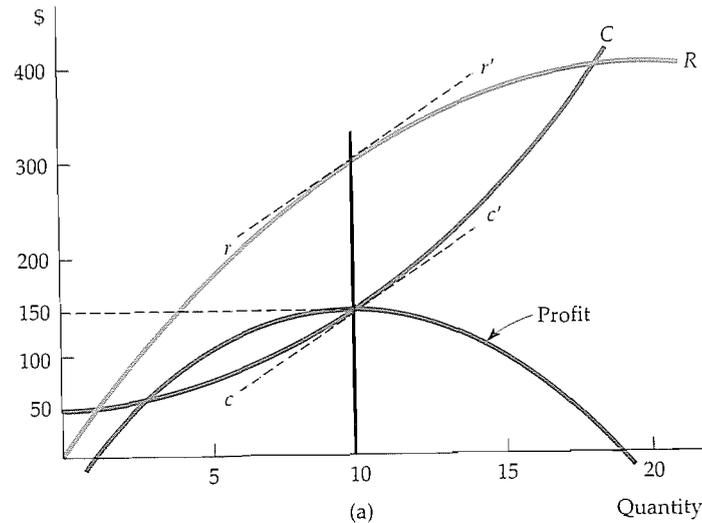


FIGURE 10.3 Example of Profit Maximization

Part (a) shows total revenue R , total cost C , and profit, the difference between the two. Part (b) shows average and marginal revenue and average and marginal cost. Marginal revenue is the slope of the total revenue curve, and marginal cost is the slope of the total cost curve. The profit-maximizing output is $Q^* = 10$, the point where marginal revenue equals marginal cost. At this output level, the slope of the profit curve is zero, and the slopes of the total revenue and total cost curves are equal. The profit per unit is \$15, the difference between average revenue and average cost. Because 10 units are produced, total profit is \$150.

The slope of the revenue curve is $\Delta R/\Delta Q$, or marginal revenue, and the slope of the cost curve is $\Delta C/\Delta Q$, or marginal cost. Because profit is maximized when marginal revenue equals marginal cost, the slopes are equal.

Figure 10.3(b) shows both the corresponding average and marginal revenue curves and average and marginal cost curves. Marginal revenue and marginal cost intersect at $Q^* = 10$. At this quantity, average cost is \$15 per unit and price is \$30 per unit. Thus average profit is $\$30 - \$15 = \$15$ per unit. Because 10 units are sold, profit is $(10)(\$15) = \150 , the area of the shaded rectangle.

A Rule of Thumb for Pricing

We know that price and output should be chosen so that marginal revenue equals marginal cost, but how can the manager of a firm find the correct price and output level in practice? Most managers have only limited knowledge of the average and marginal revenue curves that their firms face. Similarly, they might know the firm's marginal cost only over a limited output range. We therefore want to translate the condition that marginal revenue should equal marginal cost into a rule of thumb that can be more easily applied in practice.

To do this, we first rewrite the expression for marginal revenue:

$$MR = \frac{\Delta R}{\Delta Q} = \frac{\Delta(PQ)}{\Delta Q}$$

Note that the extra revenue from an incremental unit of quantity, $\Delta(PQ)/\Delta Q$, has two components:

1. Producing one extra unit and selling it at price P brings in revenue $(1)(P) = P$.
2. But because the firm faces a downward-sloping demand curve, producing and selling this extra unit also results in a small drop in price $\Delta P/\Delta Q$, which reduces the revenue from all units sold (i.e., a change in revenue $Q[\Delta P/\Delta Q]$).

Thus,

$$MR = P + Q \frac{\Delta P}{\Delta Q} = P + P \left(\frac{Q}{P} \right) \left(\frac{\Delta P}{\Delta Q} \right)$$

We obtained the expression on the right by taking the term $Q(\Delta P/\Delta Q)$ and multiplying and dividing it by P . Recall that the elasticity of demand is defined as $E_d = (P/Q)(\Delta Q/\Delta P)$. Thus $(Q/P)(\Delta P/\Delta Q)$ is the reciprocal of the elasticity of demand, $1/E_d$, measured at the profit-maximizing output, and

$$MR = P + P(1/E_d)$$

Now, because the firm's objective is to maximize profit, we can set marginal revenue equal to marginal cost:

$$P + P(1/E_d) = MC$$

which can be rearranged to give us

$$\frac{P - MC}{P} = -\frac{1}{E_d} \quad (10.1)$$

This relationship provides a rule of thumb for pricing. The left-hand side, $(P - MC)/P$, is the markup over marginal cost as a percentage of price. The relationship says that this markup should equal minus the inverse of the elasticity of demand.⁴ (This figure will be a positive number because the elasticity of demand is negative.) Equivalently, we can rearrange this equation to express price directly as a markup over marginal cost:

$$P = \frac{MC}{1 + (1/E_d)} \quad (10.2)$$

For example, if the elasticity of demand is -4 and marginal cost is \$9 per unit, price should be $\$9/(1 - 1/4) = \$9/.75 = \$12$ per unit.

How does the price set by a monopolist compare with the price under competition? In Chapter 8, we saw that in a perfectly competitive market, price equals marginal cost. A monopolist charges a price that exceeds marginal cost, but by an amount that depends inversely on the elasticity of demand. As the markup equation (10.1) shows, if demand is extremely elastic, E_d is a large negative number, and price will be very close to marginal cost. In that case, a monopolized market will look much like a competitive one. In fact, when demand is very elastic, there is little benefit to being a monopolist.

EXAMPLE 10.1 Astra-Merck Prices Prilosec

In 1995, a new drug developed by Astra-Merck became available for the long-term treatment of ulcers. The drug, Prilosec, represented a new generation of antiulcer medication. Other drugs to treat ulcer conditions were already on the market: Tagamet had been introduced in 1977, Zantac in 1983, Pepcid in 1986, and Axid in 1988. As we explained in Example 1.1, these four drugs worked in much the same way to reduce the stomach's secretion of acid. Prilosec, however, was based on a very different biochemical mechanism and was much more effective than these earlier drugs. By 1996, it had become the best-selling drug in the world and faced no major competitor.⁵

⁴ Remember that this markup equation applies at the point of a profit maximum. If both the elasticity of demand and marginal cost vary considerably over the range of outputs under consideration, you may have to know the entire demand and marginal cost curves to determine the optimum output level. On the other hand, you can use this equation to check whether a particular output level and price are optimal.

⁵ Prilosec, developed through a joint venture of the Swedish firm Astra and the U.S. firm Merck, was introduced in 1989, but only for the treatment of gastroesophageal reflux disease, and was approved for short-term ulcer treatment in 1991. It was the approval for long-term ulcer treatment in 1995, however, that created a very large market for the drug. In 1998, Astra bought Merck's share of the rights to Prilosec. In 1999, Astra acquired the firm Zeneca and is now called AstraZeneca.

In 1995, Astra-Merck was pricing Prilosec at about \$3.50 per daily dose. (By contrast, the prices for Tagamet and Zantac were about \$1.50 to \$2.25 per daily dose.) Is this pricing consistent with the markup formula (10.2)? The marginal cost of producing and packaging Prilosec is only about 30 to 40 cents per daily dose. This low marginal cost implies that the price elasticity of demand, E_d , should be in the range of roughly -1.0 to -1.2 . Based on statistical studies of pharmaceutical demands, this is indeed a reasonable estimate for the demand elasticity. Thus, setting the price of Prilosec at a markup exceeding 400 percent over marginal cost is consistent with our rule of thumb for pricing.

Shifts in Demand

In a competitive market, there is a clear relationship between price and the quantity supplied. That relationship is the supply curve, which, as we saw in Chapter 8, represents the marginal cost of production for the industry as a whole. The supply curve tells us how much will be produced at every price.

A monopolistic market has no supply curve. In other words, there is no one-to-one relationship between price and the quantity produced. The reason is that the monopolist's output decision depends not only on marginal cost but also on the shape of the demand curve. As a result, shifts in demand do not trace out the series of prices and quantities that correspond to a competitive supply curve. Instead, shifts in demand can lead to changes in price with no change in output, changes in output with no change in price, or changes in both.

This principle is illustrated in Figure 10.4(a) and (b). In both parts of the figure, the demand curve is initially D_1 , the corresponding marginal revenue curve is MR_1 , and the monopolist's initial price and quantity are P_1 and Q_1 . In Figure 10.4(a), the demand curve is shifted down and rotated. The new demand and marginal revenue curves are shown as D_2 and MR_2 . Note that MR_2 intersects the marginal cost curve at the same point that MR_1 does. As a result, the quantity produced stays the same. Price, however, falls to P_2 .

In Figure 10.4(b), the demand curve is shifted up and rotated. The new marginal revenue curve MR_2 intersects the marginal cost curve at a larger quantity, Q_2 instead of Q_1 . But the shift in the demand curve is such that the price charged is exactly the same.

Shifts in demand usually cause changes in both price and quantity. But the special cases shown in Figure 10.4 illustrate an important distinction between monopoly and competitive supply. A competitive industry supplies a specific quantity at every price. No such relationship exists for a monopolist, which, depending on how demand shifts, might supply several different quantities at the same price, or the same quantity at different prices.

The Effect of a Tax

A tax on output can also have a different effect on a monopolist than on a competitive industry. In Chapter 9, we saw that when a specific (i.e., per-unit) tax is imposed on a competitive industry, the market price rises by an amount that is less than the tax, and that the burden of the tax is shared by producers and consumers. Under monopoly, however, price can sometimes rise by *more* than the amount of the tax.

In §9.6, we explain that a specific tax is a tax of a certain amount of money per unit sold, and we show how the tax affects price and quantity.

In §8.2, we explain that a perfectly competitive firm will choose its output so that marginal cost equals price.

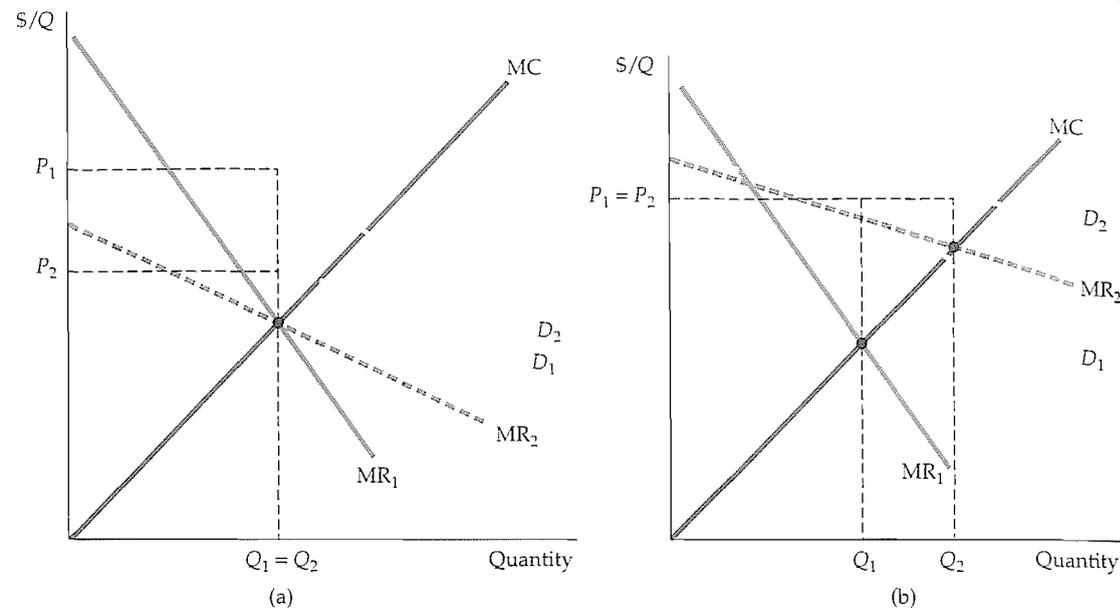


FIGURE 10.4 Shifts in Demand

Shifting the demand curve shows that a monopolistic market has no supply curve—i.e., there is no one-to-one relationship between price and quantity produced. In (a), the demand curve D_1 shifts to new demand curve D_2 . But the new marginal revenue curve MR_2 intersects marginal cost at the same point that the old marginal revenue curve MR_1 did. The profit-maximizing output therefore remains the same, although price falls from P_1 to P_2 . In (b), the new marginal revenue curve MR_2 intersects marginal cost at a higher output level Q_2 . But because demand is now more elastic, price remains the same.

Analyzing the effect of a tax on a monopolist is straightforward. Suppose a specific tax of t dollars per unit is levied, so that the monopolist must remit t dollars to the government for every unit it sells. Therefore, the firm's marginal (and average) cost is increased by the amount of the tax t . If MC was the firm's original marginal cost, its optimal production decision is now given by

$$MR = MC + t$$

Graphically, we shift the marginal cost curve upward by an amount t , and find the new intersection with marginal revenue. Figure 10.5 shows this. Here Q_0 and P_0 are the quantity and price before the tax is imposed, and Q_1 and P_1 are the quantity and price after the tax.

Shifting the marginal cost curve upward results in a smaller quantity and higher price. Sometimes price increases by less than the tax, but not always—in Figure 10.5, price increases by *more* than the tax. This would be impossible in a competitive market, but it can happen with a monopolist because the relationship between price and marginal cost depends on the elasticity of demand. Suppose, for example, that a monopolist faces a constant elasticity demand curve, with elasticity -2 , and has constant marginal cost MC . Equation (10.2) then tells us that price will equal twice marginal cost. With a tax t , marginal cost increases to $MC + t$, so price increases to $2(MC + t) = 2MC + 2t$; that is, it rises by twice the amount of the tax. (However, the monopolist's profit nonetheless falls with the tax.)

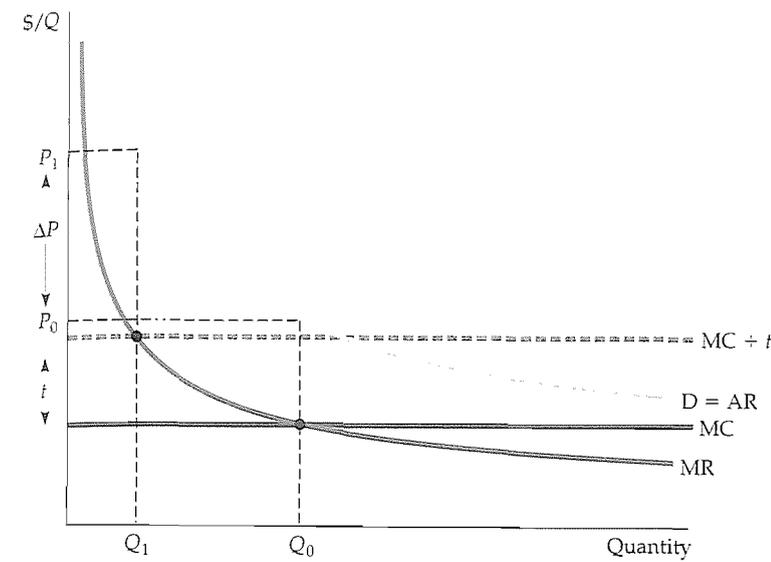


FIGURE 10.5 Effect of Excise Tax on Monopolist

With a tax t per unit, the firm's effective marginal cost is increased by the amount t to $MC + t$. In this example, the increase in price ΔP is larger than the tax t .

*The Multiplant Firm

We have seen that a firm maximizes profit by setting output at a level where marginal revenue equals marginal cost. For many firms, production takes place in two or more different plants whose operating costs can differ. However, the logic used in choosing output levels is very similar to that for the single-plant firm.

Suppose a firm has two plants. What should its total output be, and how much of that output should each plant produce? We can find the answer intuitively in two steps.

- **Step 1.** Whatever the total output, it should be divided between the two plants so that *marginal cost is the same in each plant*. Otherwise, the firm could reduce its costs and increase its profit by reallocating production. For example, if marginal cost at Plant 1 were higher than at Plant 2, the firm could produce the same output at a lower total cost by producing less at Plant 1 and more at Plant 2.
- **Step 2.** We know that total output must be such that *marginal revenue equals marginal cost*. Otherwise, the firm could increase its profit by raising or lowering total output. For example, suppose marginal costs were the same at each plant, but marginal revenue exceeded marginal cost. In that case, the firm would do better by producing more at both plants because the revenue earned from the additional units would exceed the cost. Since marginal costs must be the same at each plant, and marginal revenue must equal marginal cost, we see that profit is maximized when *marginal revenue equals marginal cost at each plant*.

We can also derive this result algebraically. Let Q_1 and C_1 be the output and cost of production for Plant 1, Q_2 and C_2 be the output and cost of production for Plant 2, and $Q_T = Q_1 + Q_2$ be total output. Then profit is

$$\pi = PQ_T - C_1(Q_1) - C_2(Q_2)$$

In §8.2, we explain that a firm maximizes its profit by choosing the output at which marginal revenue is equal to marginal cost.

The firm should increase output from each plant until the incremental profit from the last unit produced is zero. Start by setting incremental profit from output at Plant 1 to zero:

$$\frac{\Delta \pi}{\Delta Q_1} = \frac{\Delta(PQ_T)}{\Delta Q_1} - \frac{\Delta C_1}{\Delta Q_1} = 0$$

Here $\Delta(PQ_T)/\Delta Q_1$ is the revenue from producing and selling one more unit, i.e., marginal revenue, MR, for all of the firm's output. The next term, $\Delta C_1/\Delta Q_1$, is marginal cost at Plant 1, MC_1 . We thus have $MR - MC_1 = 0$, or

$$MR = MC_1$$

Similarly, we can set incremental profit from output at Plant 2 to zero,

$$MR = MC_2$$

Putting these relations together, we see that the firm should produce so that

$$MR = MC_1 = MC_2 \quad (10.3)$$

Figure 10.6 illustrates this principle for a firm with two plants. MC_1 and MC_2 are the marginal cost curves for the two plants. (Note that Plant 1 has higher marginal costs than Plant 2.) Also shown is a curve labeled MC_T . This is the firm's total marginal cost and is obtained by horizontally summing MC_1 and

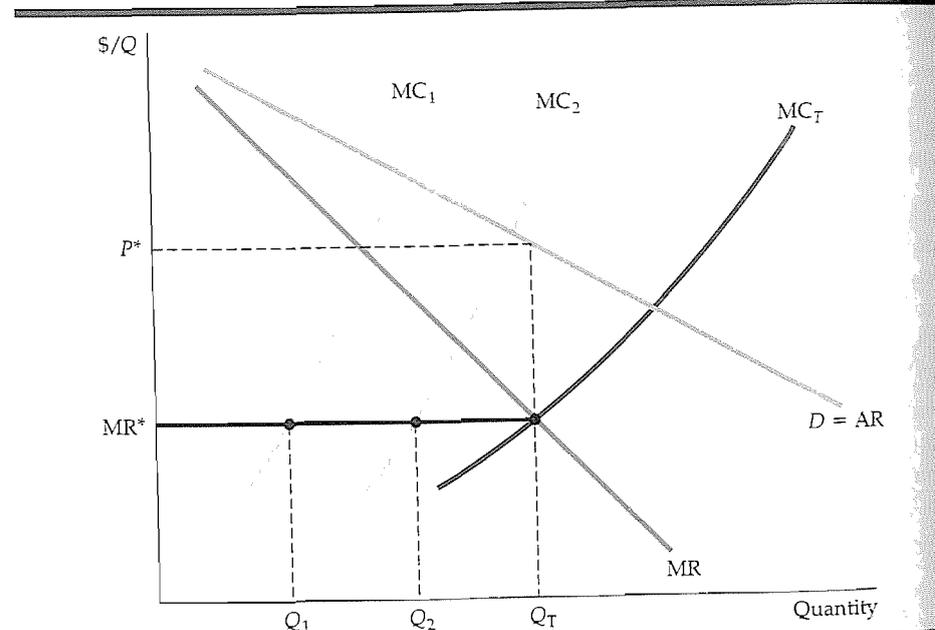


FIGURE 10.6 Production with Two Plants

A firm with two plants maximizes profits by choosing output levels Q_1 and Q_2 so that marginal revenue MR (which depends on total output) equals marginal costs for each plant, MC_1 and MC_2 .

MC_2 .⁶ Now we can find the profit-maximizing output levels Q_1 , Q_2 , and Q_T . First, find the intersection of MC_T with MR ; that point determines total output Q_T . Next, draw a horizontal line from that point on the marginal revenue curve to the vertical axis; point MR^* determines the firm's marginal revenue. The intersections of the marginal revenue line with MC_1 and MC_2 give the outputs Q_1 and Q_2 for the two plants, as in equation (10.3).

Note that total output Q_T determines the firm's marginal revenue (and hence its price P^*). Q_1 and Q_2 , however, determine marginal costs at each of the two plants. Because MC_T was found by horizontally summing MC_1 and MC_2 , we know that $Q_1 + Q_2 = Q_T$. Thus these output levels satisfy the condition that $MR = MC_1 = MC_2$.

10.2 Monopoly Power

Pure monopoly is rare. Markets in which several firms compete with one another are much more common. We say more about the forms this competition can take in Chapters 12 and 13. But we should explain here why each firm in a market with several firms is likely to face a downward-sloping demand curve, and, as a result, produce so that price exceeds marginal cost.

Suppose, for example, that four firms produce toothbrushes, which have the market demand curve shown in Figure 10.7(a). Let's assume that these

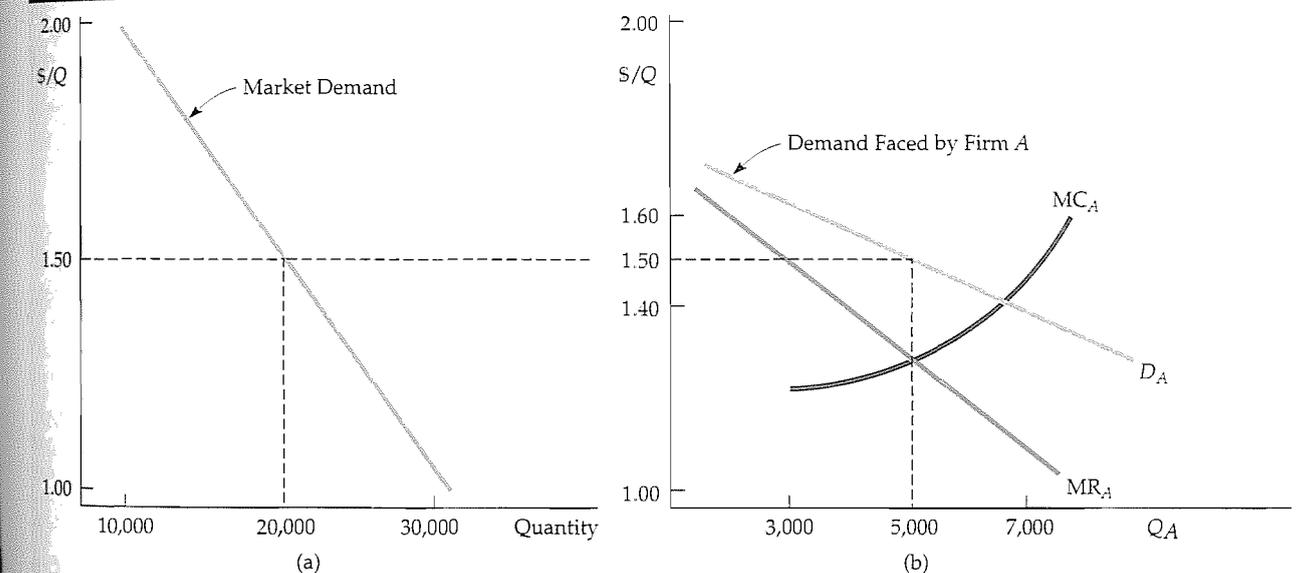


FIGURE 10.7 The Demand for Toothbrushes

Part (a) shows the market demand for toothbrushes. Part (b) shows the demand for toothbrushes as seen by Firm A. At a market price of \$1.50, elasticity of market demand is -1.5 . Firm A, however, sees a much more elastic demand curve D_A because of competition from other firms. At a price of \$1.50, Firm A's demand elasticity is -6 . Still, Firm A has some monopoly power: Its profit-maximizing price is \$1.50, which exceeds marginal cost.

⁶ Note the similarity to the way we obtained a competitive industry's supply curve in Chapter 8 by horizontally summing the marginal cost curves of the individual firms.

four firms are producing an aggregate of 20,000 toothbrushes per day (5000 per day each) and selling them at \$1.50 each. Note that market demand is relatively inelastic; you can verify that at this \$1.50 price, the elasticity of demand is -1.5 .

Now suppose that Firm A is deciding whether to lower its price to increase sales. To make this decision, it needs to know how its sales would respond to a change in its price. In other words, it needs some idea of the demand curve it faces, as opposed to the *market* demand curve. A reasonable possibility is shown in Figure 10.7(b), where the firm's demand curve D_A is much more elastic than the market demand curve. (At the \$1.50 price the elasticity is -6.0 .) The firm might predict that by raising price from \$1.50 to \$1.60, its sales will drop—say, from 5000 units to 3000—as consumers buy more toothbrushes from other firms. (If *all* firms raised their prices to \$1.60, sales for Firm A would fall only to 4500.) But for several reasons, sales won't drop to zero as they would in a perfectly competitive market. First, if Firm A's toothbrushes are a little different from its competitors, some consumers will pay a bit more for them. Second, other firms might also raise their prices. Similarly, Firm A might anticipate that by lowering its price from \$1.50 to \$1.40, it can sell more, perhaps 7000 toothbrushes instead of 5000. But it will not capture the entire market. Some consumers might still prefer the competitors' toothbrushes, and the competitors might also lower their prices.

Thus Firm A's demand curve depends both on how much its product differs from its competitors' products and on how the four firms compete with one another. We will discuss product differentiation and interfirm competition in Chapters 12 and 13. But one important point should be clear: *Firm A is likely to face a demand curve which is more elastic than the market demand curve, but which is not infinitely elastic like the demand curve facing a perfectly competitive firm.*

Given knowledge of its demand curve, how much should Firm A produce? The same principle applies: The profit-maximizing quantity equates marginal revenue and marginal cost. In Figure 10.7(b), that quantity is 5000 units. The corresponding price is \$1.50, which exceeds marginal cost. Thus although Firm A is not a pure monopolist, *it does have monopoly power*—it can profitably charge a price greater than marginal cost. Of course, its monopoly power is less than it would be if it had driven away the competition and monopolized the market, but it might still be substantial.

This raises two questions.

1. How can we *measure* monopoly power in order to compare one firm with another? (So far we have been talking about monopoly power only in *qualitative* terms.)
2. What are the *sources* of monopoly power, and why do some firms have more monopoly power than others?

We address both these questions below, although a more complete answer to the second question will be provided in Chapters 12 and 13.

Measuring Monopoly Power

Remember the important distinction between a perfectly competitive firm and a firm with monopoly power: *For the competitive firm, price equals marginal cost; for the firm with monopoly power, price exceeds marginal cost.* Therefore, a natural way to measure monopoly power is to examine the extent to which the profit-maxi-

mizing price exceeds marginal cost. In particular, we can use the markup ratio of price minus marginal cost to price that we introduced earlier as part of a rule of thumb for pricing. This measure of monopoly power, introduced by economist Abba Lerner in 1934, is called the **Lerner Index of Monopoly Power**. It is the difference between price and marginal cost, divided by price. Mathematically:

$$L = (P - MC)/P$$

The Lerner index always has a value between zero and one. For a perfectly competitive firm, $P = MC$, so that $L = 0$. The larger L , the greater the degree of monopoly power.

This index of monopoly power can also be expressed in terms of the elasticity of demand facing the firm. Using equation (10.1), we know that

$$L = (P - MC)/P = -1/E_d \quad (10.4)$$

Remember, however, that E_d is now the elasticity of the *firm's* demand curve, not the market demand curve. In the toothbrush example discussed above, the elasticity of demand for Firm A is -6.0 , and the degree of monopoly power is $1/6 = 0.167$.

Note that considerable monopoly power does not necessarily imply high profits. Profit depends on *average* cost relative to price. Firm A might have more monopoly power than Firm B but earn a lower profit because of higher average costs.

The Rule of Thumb for Pricing

In the previous section, we used equation (10.2) to compute price as a simple markup over marginal cost:

$$P = \frac{MC}{1 + (1/E_d)}$$

This relationship provides a rule of thumb for *any* firm with monopoly power. We must remember, however, that E_d is the elasticity of demand for the *firm*, not the elasticity of *market* demand.

It is harder to determine the elasticity of demand for the firm than for the market because the firm must consider how its competitors will react to price changes. Essentially, the manager must estimate the percentage change in the firm's unit sales that is likely to result from a 1-percent change in the firm's price. This estimate might be based on a formal model or on the manager's intuition and experience.

Given an estimate of the firm's elasticity of demand, the manager can calculate the proper markup. If the firm's elasticity of demand is large, this markup

Lerner Index of Monopoly Power Measure of monopoly power calculated as excess of price over marginal cost as a fraction of price.

⁷ There are three problems with applying the Lerner index to the analysis of public policy toward firms. First, because marginal cost is difficult to measure, average variable cost is often used in Lerner index calculations. Second, if the firm prices below its optimal price (possibly to avoid legal scrutiny), its potential monopoly power will not be noted by the index. Third, the index ignores dynamic aspects of pricing such as effects of the learning curve and shifts in demand. See Robert S. Pindyck, "The Measurement of Monopoly Power in Dynamic Markets," *Journal of Law and Economics* 28 (April 1985): 193–222.

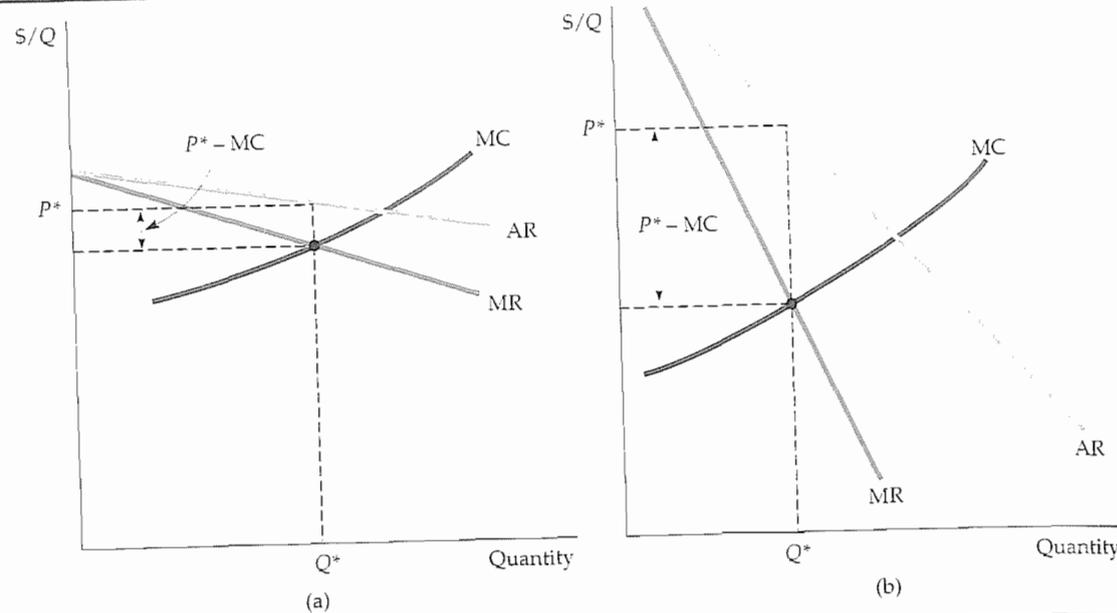


FIGURE 10.8 Elasticity of Demand and Price Markup

The markup $(P - MC)/P$ is equal to minus the inverse of the elasticity of demand facing the firm. If the firm's demand is elastic as in (a), the markup is small and the firm has little monopoly power. The opposite is true if demand is relatively inelastic, as in (b).

will be small (and we can say that the firm has very little monopoly power). If the firm's elasticity of demand is small, this markup will be large (and the firm will have considerable monopoly power). Figures 10.8(a) and 10.8(b) illustrate these two extremes.

EXAMPLE 10.2 Markup Pricing: Supermarkets to Designer Jeans

Three examples should help clarify the use of markup pricing. Consider a retail supermarket chain. Although the elasticity of market demand for food is small (about -1), several supermarkets usually serve most areas, so no single supermarket can raise its prices very much without losing many customers to other stores. As a result, the elasticity of demand for any one supermarket is often as large as -10 . Substituting this number for E_d in equation (10.2), we find $P = MC/(1 - 0.1) = MC/(0.9) = (1.11)MC$. In other words, the manager of a typical supermarket should set prices about 11 percent above marginal cost. For a reasonably wide range of output levels (over which the size of the store and the number of its employees will remain fixed), marginal cost includes the cost of purchasing the food at wholesale, plus the costs of storing the food, arranging it on the shelves, etc. For most supermarkets, the markup is indeed about 10 or 11 percent.

Small convenience stores, which are often open 7 days a week and even 24 hours a day, typically charge higher prices than supermarkets. Why? Because a convenience store faces a less elastic demand curve. Its customers are generally less price sensitive. They might need a quart of milk or a loaf of bread

late at night, or may find it inconvenient to drive to the supermarket. The elasticity of demand for a convenience store is about -5 , so the markup equation implies that its prices should be about 25 percent above marginal cost, as indeed they typically are.

The Lerner index, $(P - MC)/P$, tells us that the convenience store has more monopoly power, but does it make larger profits? No. Because its volume is far smaller and its average fixed costs are larger, it usually earns a much smaller profit than a large supermarket, despite its higher markup.

Finally, consider a producer of designer jeans. Many companies produce jeans, but some consumers will pay much more for jeans with a designer label. Just how much more they will pay—or more exactly, how much sales will drop in response to higher prices—is a question that the producer must carefully consider because it is critical in determining the price at which the clothing will be sold (at wholesale to retail stores, which then mark up the price further). With designer jeans, demand elasticities in the range of -3 to -4 are typical for the major labels. This means that price should be 33 to 50 percent higher than marginal cost. Marginal cost is typically \$12 to \$18 per pair, and the wholesale price is in the \$18 to \$27 range.

EXAMPLE 10.3 The Pricing of Prerecorded Videocassettes

During the mid-1980s, the number of households owning videocassette recorders (VCRs) grew rapidly, as did the markets for rentals and sales of prerecorded cassettes. Although many more videocassettes are rented through small retail outlets than are sold outright, the market for sales is large and growing. Producers, however, found it difficult to decide what price to charge for cassettes. As a result, in 1985 popular movies were selling for vastly different prices, as the data for that year show in Table 10.2.

Note that while *The Empire Strikes Back* was selling for nearly \$80, *Star Trek*, a film that appealed to the same audience and was about as popular, sold for only about \$25. These price differences reflected uncertainty and a wide

TABLE 10.2 Prices of Videos in 1985 and 1999

1985		1999	
TITLE	RETAIL PRICE (\$)	TITLE	RETAIL PRICE (\$)
Purple Rain	\$29.98	Austin Powers	\$10.49
Raiders of the Lost Ark	24.95	A Bug's Life	17.99
Jane Fonda Workout	59.95	There's Something about Mary	13.99
The Empire Strikes Back	79.98	Tae-Bo Workout	24.47
An Officer and a Gentleman	24.95	Lethal Weapon 4	16.99
Star Trek: The Motion Picture	24.95	Men in Black	12.99
Star Wars	39.98	Armageddon	15.86

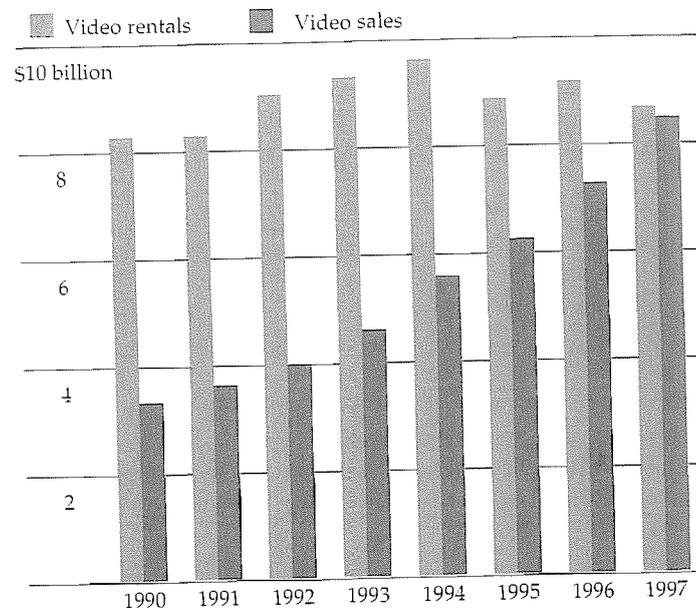


FIGURE 10.9 Video Rentals and Sales

Between 1990 and 1997, lower prices induced consumers to buy many more videos. While revenues from sales more than doubled, rental revenues were flat.

divergence of views on pricing by producers. The issue was whether lower prices would induce consumers to buy videocassettes rather than rent them. Because producers do not share in the retailers' revenues from rentals, they should charge a low price for cassettes only if that will induce enough consumers to buy them. Because the market was young, producers had no good estimates of the elasticity of demand, so they based prices on hunches or trial and error.⁸

As the market matured, however, sales data and market research studies put pricing decisions on firmer ground. They strongly indicated that demand was elastic and that the profit-maximizing price was in the range of \$15 to \$30. As one industry analyst said, "People are becoming collectors. . . . As you lower the price you attract households that would not have considered buying at a higher price point."⁹ By the 1990s, most producers had lowered prices across the board. As Table 10.2 shows, in 1999 prices of top-selling videos were considerably lower than in 1985. As a result of these price declines, sales of videos increased steadily during the 1990s, as did profits from these sales. As Figure 10.9 shows, revenues from video sales more than doubled from 1990 to 1997, while revenues from rentals remained fairly flat.

⁸ "Video Producers Debate the Value of Price Cuts," *New York Times*, February 19, 1985.

⁹ "Studios Now Stressing Video Sales Over Rentals," *New York Times*, October 17, 1989. For a detailed study of videocassette pricing, see Carl E. Enomoto and Soumendra N. Ghosh, "Pricing in the Home-Video Market" (working paper, New Mexico State University, 1992).

10.3 Sources of Monopoly Power

Why do some firms have considerable monopoly power while other firms have little or none? Remember that monopoly power is the ability to set price above marginal cost and that the amount by which price exceeds marginal cost depends inversely on the elasticity of demand facing the firm. As equation (10.3) shows, the less elastic its demand curve, the more monopoly power a firm has. The ultimate determinant of monopoly power is therefore the firm's elasticity of demand. Thus we should rephrase our question: Why do some firms (e.g., a supermarket chain) face demand curves that are more elastic than those faced by others (e.g., a producer of designer clothing)?

Three factors determine a firm's elasticity of demand:

1. *The elasticity of market demand.* Because the firm's own demand will be at least as elastic as market demand, the elasticity of market demand limits the potential for monopoly power.
2. *The number of firms in the market.* If there are many firms, it is unlikely that any one firm will be able to affect price significantly.
3. *The interaction among firms.* Even if only two or three firms are in the market, each firm will be unable to profitably raise price very much if the rivalry among them is aggressive, with each firm trying to capture as much of the market as it can. Let's examine each of these three determinants of monopoly power.

The Elasticity of Market Demand

If there is only one firm—a pure monopolist—its demand curve is the market demand curve. In this case, the firm's degree of monopoly power depends completely on the elasticity of market demand. More often, however, several firms compete with one another; then the elasticity of market demand sets a lower limit on the magnitude of the elasticity of demand for each firm. Recall our example of the toothbrush producers illustrated in Figure 10.7. The market demand for toothbrushes might not be very elastic, but each firm's demand will be more elastic. (In Figure 10.7, the elasticity of market demand is -1.5 , and the elasticity of demand for each firm is -6 .) A particular firm's elasticity depends on how the firms compete with one another. But no matter how they compete, the elasticity of demand for each firm could never become smaller in magnitude than -1.5 .

Because the demand for oil is fairly inelastic (at least in the short run), OPEC could raise oil prices far above marginal production cost during the 1970s and early 1980s. Because the demands for such commodities as coffee, cocoa, tin, and copper are much more elastic, attempts by producers to cartelize these markets and raise prices have largely failed. In each case, the elasticity of market demand limits the potential monopoly power of individual producers.

The Number of Firms

The second determinant of a firm's demand curve—and thus of its monopoly power—is the number of firms in its market. Other things being equal, the monopoly power of each firm will fall as the number of firms increases. As more and more firms compete, each firm will find it harder to raise prices and avoid losing sales to other firms.

What matters, of course, is not just the total number of firms, but the number of "major players"—firms with significant market share. For example, if only two large firms account for 90 percent of sales in a market, with another 20 firms accounting for the remaining 10 percent, the two large firms might have considerable monopoly power. When only a few firms account for most of the sales in a market, we say that the market is highly *concentrated*.¹⁰

It is sometimes said (not always jokingly) that the greatest fear of American business is competition. That may or may not be true. But we would certainly expect that when only a few firms are in a market, their managers will prefer that no new firms enter. An increase in the number of firms can only reduce the monopoly power of each incumbent firm. An important aspect of competitive strategy (discussed in detail in Chapter 13) is finding ways to create **barriers to entry**—conditions that deter entry by new competitors.

barrier to entry Condition that impedes entry by new competitors.

Sometimes there are natural barriers to entry. For example, one firm may have a *patent* on the technology needed to produce a particular product. This makes it impossible for other firms to enter the market, at least until the patent expires. Other legally created rights work in the same way—a *copyright* can limit the sale of a book, music, or a computer software program to a single company, and the need for a government *license* can prevent new firms from entering the markets for telephone service, television broadcasting, or interstate trucking. Finally, *economies of scale* may make it too costly for more than a few firms to supply the entire market. In some cases, economies of scale may be so large that it is most efficient for a single firm—a *natural monopoly*—to supply the entire market. We will discuss scale economies and natural monopoly in more detail shortly.

In §7.4, we explain that a firm enjoys economies of scale when it can double its output with less than a doubling of cost.

The Interaction Among Firms

The ways in which competing firms interact is also an important—and sometimes the most important—determinant of monopoly power. Suppose there are four firms in a market. They might compete aggressively, undercutting one another's prices to capture more market share. This could drive prices down to nearly competitive levels. Each firm will fear that if it raises its price it will be undercut and lose market share. As a result, it will have little monopoly power.

On the other hand, the firms might not compete much. They might even collude (in violation of the antitrust laws), agreeing to limit output and raise prices. Raising prices in concert rather than individually is more likely to be profitable, so collusion can generate substantial monopoly power.

We will discuss the interaction among firms in detail in Chapters 12 and 13. Now we simply want to point out that other things being equal, monopoly power is smaller when firms compete aggressively and is larger when they cooperate.

Remember that a firm's monopoly power often changes over time, as its operating conditions (market demand and cost), its behavior, and the behavior of its competitors change. Monopoly power must therefore be thought of in a dynamic context. For example, the market demand curve might be very inelastic in the short run but much more elastic in the long run. (Because this is the case with oil, the OPEC cartel enjoyed considerable short-run but much less long-run monopoly power.) Furthermore, real or potential monopoly power in the short

¹⁰ A statistic called the *concentration ratio*, which measures the fraction of sales accounted for by, say, the four largest firms, is often used to describe the concentration of a market. Concentration is one, but not the only, determinant of market power.

run can make an industry more competitive in the long run. Large short-run profits can induce new firms to enter an industry, thereby reducing monopoly power over the longer term.

10.4 The Social Costs of Monopoly Power

In a competitive market, price equals marginal cost. Monopoly power, on the other hand, implies that price exceeds marginal cost. Because monopoly power results in higher prices and lower quantities produced, we would expect it to make consumers worse off and the firm better off. But suppose we value the welfare of consumers the same as that of producers. In the aggregate, does monopoly power make consumers and producers better or worse off?

We can answer this question by comparing the consumer and producer surplus that results when a competitive industry produces a good with the surplus that results when a monopolist supplies the entire market.¹¹ (We assume that the competitive market and the monopolist have the same cost curves.) Figure 10.10 shows the average and marginal revenue curves and marginal cost curve for the monopolist. To maximize profit, the firm produces at the point where marginal revenue equals marginal cost, so that the price and quantity are P_m and Q_m . In a competitive market, price must equal marginal cost, so the competitive price and quantity, P_c and Q_c , are found at the intersection of the average revenue

In §9.1, we explain that consumer surplus is the total benefit or value that consumers receive beyond what they pay for a good; producer surplus is the analogous measure for producers.

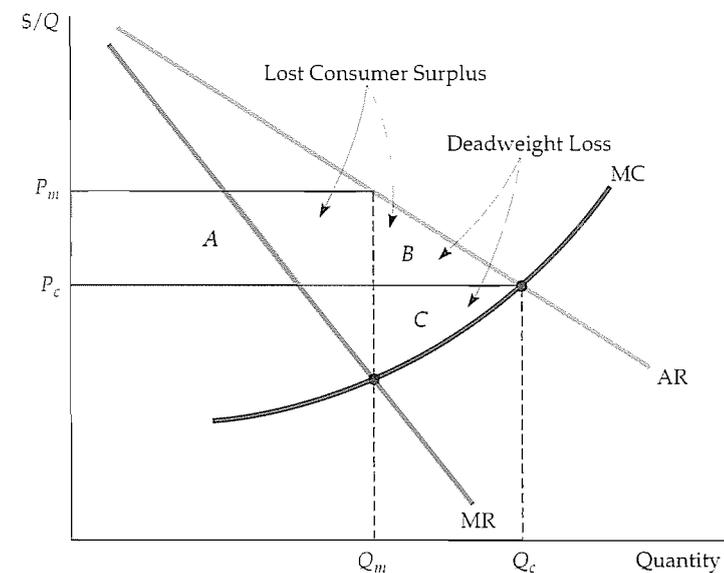


FIGURE 10.10 Deadweight Loss from Monopoly Power

The shaded rectangle and triangles show changes in consumer and producer surplus when moving from competitive price and quantity, P_c and Q_c , to a monopolist's price and quantity, P_m and Q_m . Because of the higher price, consumers lose $A + B$ and producer gains $A - C$. The deadweight loss is $-B - C$.

¹¹ If there were two or more firms, each with some monopoly power, the analysis would be more complex. However, the basic results would be the same.

(demand) curve and the marginal cost curve. Now let's examine how surplus changes if we move from the competitive price and quantity, P_c and Q_c , to the monopoly price and quantity, P_m and Q_m .

Under monopoly, the price is higher and consumers buy less. Because of the higher price, those consumers who buy the good lose surplus of an amount given by rectangle A . Those consumers who do not buy the good at price P_m but who will buy at price P_c also lose surplus—namely, an amount given by triangle B . The total loss of consumer surplus is therefore $A + B$. The producer, however, gains rectangle A by selling at the higher price but loses triangle C , the additional profit it would have earned by selling $Q_c - Q_m$ at price P_c . The total gain in producer surplus is therefore $A - C$. Subtracting the loss of consumer surplus from the gain in producer surplus, we see a net loss of surplus given by $B + C$. This is the *deadweight loss from monopoly power*. Even if the monopolist's profits were taxed away and redistributed to the consumers of its products, there would be an inefficiency because output would be lower than under conditions of competition. The deadweight loss is the social cost of this inefficiency.

Rent Seeking

In practice, the social cost of monopoly power is likely to exceed the deadweight loss in triangles B and C of Figure 10.10. The reason is that the firm may engage in **rent seeking**: spending large amounts of money in socially unproductive efforts to acquire, maintain, or exercise its monopoly power. Rent seeking might involve lobbying activities (and perhaps campaign contributions) to obtain government regulations that make entry by potential competitors more difficult. Rent-seeking activity could also involve advertising and legal efforts to avoid antitrust scrutiny. It might also mean installing but not utilizing extra production capacity to convince potential competitors that they cannot sell enough to make entry worthwhile. We would expect the economic incentive to incur rent-seeking costs to bear a direct relation to the gains from monopoly power (i.e., rectangle A minus triangle C). Therefore, the larger the transfer from consumers to the firm (rectangle A), the larger the social cost of monopoly.¹²

Here's an example. In 1996, the Archer Daniels Midland Company (ADM) successfully lobbied the Clinton administration for regulations requiring that the ethanol (ethyl alcohol) used in motor vehicle fuel be produced from corn. (The government had already planned to add ethanol to gasoline in order to reduce the country's dependence on imported oil.) Ethanol is chemically the same whether it is produced from corn, potatoes, grain, or anything else. Then why require that it be produced only from corn? Because ADM had a near monopoly on corn-based ethanol production, so the regulation would increase its gains from monopoly power.

Price Regulation

Because of its social cost, antitrust laws prevent firms from accumulating excessive amounts of monopoly power. We will say more about such laws at the end of the chapter. Here, we examine another means by which government can limit monopoly power—price regulation.

¹²The concept of rent seeking was first developed by Gordon Tullock. For more detailed discussions, see Gordon Tullock, *Rent Seeking* (Brookfield VT: Edward Elgar, 1993), or Robert D. Tollison and Roger D. Congleton, *The Economic Analysis of Rent Seeking* (Brookfield, VT: Edward Elgar, 1995).

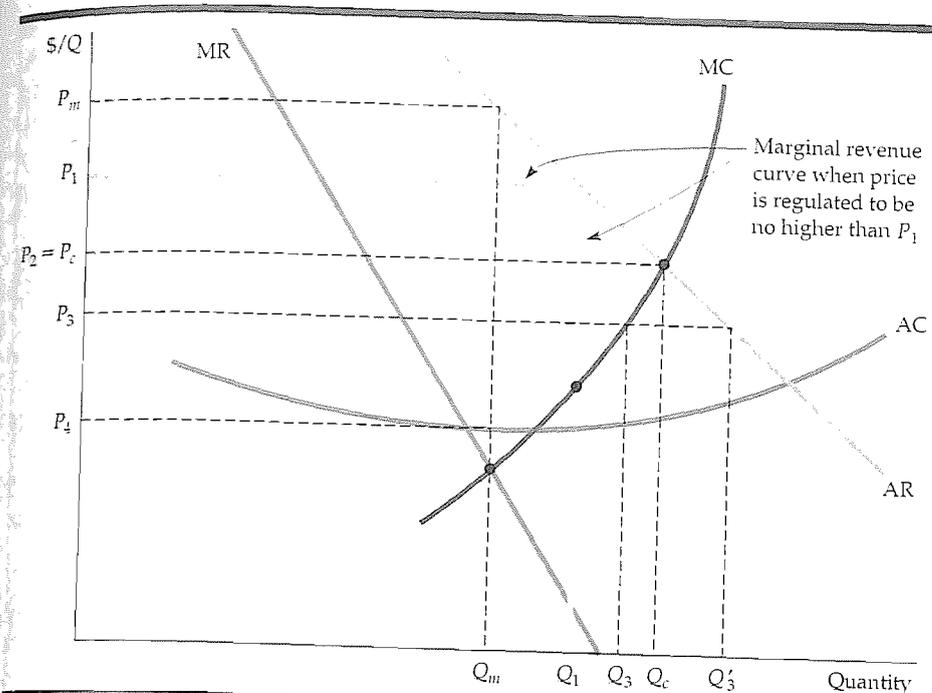


FIGURE 10.11 Price Regulation

If left alone, a monopolist produces Q_m and charges P_m . When the government imposes a price ceiling of P_1 the firm's average and marginal revenue are constant and equal to P_1 for output levels up to Q_1 . For larger output levels, the original average and marginal revenue curves apply. The new marginal revenue curve is, therefore, the dark purple line, which intersects the marginal cost curve at Q_1 . When price increases to its maximum Q_c , this is the output that would be produced by a competitive industry. Lowering price further, to P_3 , reduces output to Q_3 and causes a shortage, $Q'_3 - Q_3$.

We saw in Chapter 9 that in a competitive market, price regulation always results in a deadweight loss. This need not be the case, however, when a firm has monopoly power. On the contrary, price regulation can eliminate the deadweight loss that results from monopoly power.

Figure 10.11 illustrates price regulation. P_m and Q_m are the price and quantity that result without regulation. Now suppose the price is regulated to be no higher than P_1 . Because the firm can charge no more than P_1 for output levels up to Q_1 , its new average revenue curve is a horizontal line at P_1 . For output levels greater than Q_1 , the new average revenue curve is identical to the old average revenue curve: At these output levels, the firm will charge less than P_1 and so will be unaffected by the regulation.

The firm's new marginal revenue curve corresponds to its new average revenue curve and is shown by the dark purple line in Figure 10.10. For output levels up to Q_1 , marginal revenue equals average revenue. For output levels greater than Q_1 , the new marginal revenue curve is identical to the original curve. The firm will produce quantity Q_1 because that is the point at which its marginal revenue curve intersects its marginal cost curve. You can verify that at price P_1 and quantity Q_1 , the deadweight loss from monopoly power is reduced.

As the price is lowered further, the quantity produced continues to increase and the deadweight loss to decline. At price P_c , where average revenue and marginal cost intersect, the quantity produced has increased to the competitive level; the deadweight loss from monopoly power has been eliminated. Reducing the price even more—say, to P_3 —results in a reduction in quantity. This reduction is equivalent to imposing a price ceiling on a competitive industry. A shortage develops, $(Q'_3 - Q_3)$, in addition to the deadweight loss from regulation. As the price is lowered further, the quantity produced continues to fall and the shortage grows. Finally, if the price is lowered below P_c , the minimum average cost, the firm loses money and goes out of business.

Natural Monopoly

Price regulation is most often used for *natural monopolies*, such as local utility companies. A **natural monopoly** is a firm that can produce the entire output of the market at a cost that is lower than what it would be if there were several firms. If a firm is a natural monopoly, it is more efficient to let it serve the entire market rather than have several firms compete.

A natural monopoly usually arises when there are strong economies of scale, as illustrated in Figure 10.12. If the firm represented by the figure was broken up into two competing firms, each supplying half the market, the average cost for each would be higher than the cost incurred by the original monopoly.

Note in Figure 10.12 that because average cost is declining everywhere, marginal cost is always below average cost. If the firm were unregulated, it would produce Q_m and sell at the price P_m . Ideally, the regulatory agency would like to

push the firm's price down to the competitive level P_c . At that level, however, price would not cover average cost and the firm would go out of business. The best alternative is therefore to set the price at P_r , where average cost and average revenue intersect. In that case, the firm earns no monopoly profit, and output is as large as it can be without driving the firm out of business.

Regulation in Practice

Recall that the competitive price (P_c in Figure 10.11) is found at the point at which the firm's marginal cost and average revenue (demand) curves intersect. Likewise for a natural monopoly: The minimum feasible price (P_r in Figure 10.12) is found at the point at which average cost and demand intersect. Unfortunately, it is often difficult to determine these prices accurately in practice because the firm's demand and cost curves may shift as market conditions evolve.

Rate-of-Return Regulation As a result, the regulation of a monopoly is usually based on the rate of return that it earns on its capital. The regulatory agency determines an allowed price, so that this rate of return is in some sense "competitive" or "fair." This practice is called **rate-of-return regulation**: The maximum price allowed is based on the (expected) rate of return that the firm will earn.¹³

Unfortunately, difficult problems arise when implementing rate-of-return regulation. First, although it is a key element in determining the firm's rate of return, a firm's capital stock is difficult to value. Second, while a "fair" rate of return must be based on the firm's actual cost of capital, that cost depends in turn on the behavior of the regulatory agency (and on investors' perceptions of what future allowed rates of return will be).

The difficulty of agreeing on a set of numbers to be used in rate-of-return calculations often leads to delays in the regulatory response to changes in cost and other market conditions (not to mention long and expensive regulatory hearings). The major beneficiaries are usually lawyers, accountants, and, occasionally, economic consultants. The net result is *regulatory lag*—the delays of a year or more usually entailed in changing regulated prices.

Ironically, in the 1950s and 1960s, regulatory lag worked to the advantage of regulated firms. During those decades, costs were typically falling (usually as a result of scale economies achieved as firms grew). Thus regulatory lag allowed these firms, at least for a while, to enjoy actual rates of return greater than those ultimately deemed "fair" at the end of regulatory proceedings. Beginning in the 1970s, however, the situation changed, and regulatory lag worked to the detriment of regulated firms. For example, when oil prices rose sharply, electric utilities needed to raise their prices. Regulatory lag caused many of them to earn rates of return well below the "fair" rates they had been earning earlier.

By the 1990s, the regulatory environment in the United States had changed dramatically. Many parts of the telecommunications industry had been deregulated, as had electric utilities in many states. Because scale economies had been largely exhausted, there was no longer an argument that these firms were natural monopolies. In addition, technological change made entry by new firms relatively easy.

¹³Regulatory agencies typically use a formula like the following to determine price:

$$P = AVC + (D + T + sK)/Q$$

where AVC is average variable cost, Q is output, s is the allowed "fair" rate of return, D is depreciation, T is taxes, and K is the firm's current capital stock.

natural monopoly Firm that can produce the entire output of the market at a cost lower than what it would be if there were several firms.

rate-of-return regulation The maximum price allowed by a regulatory agency is based on the (expected) rate of return that a firm will earn.

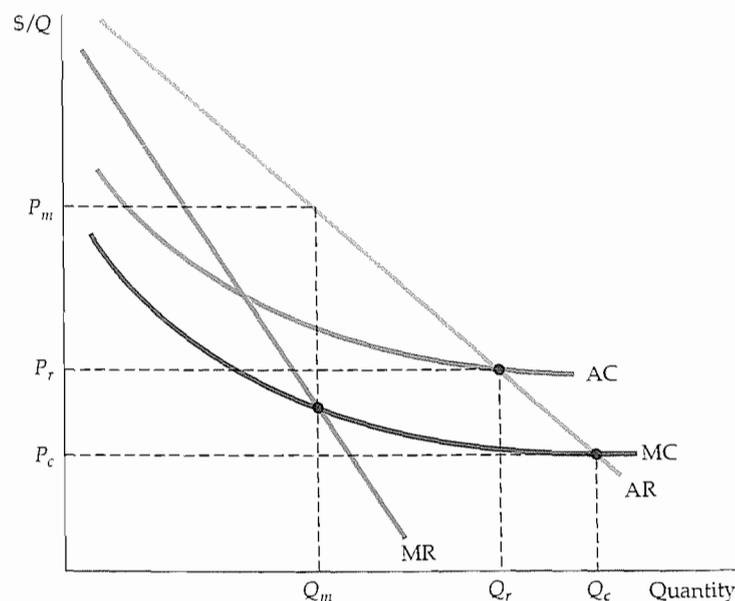


FIGURE 10.12 Regulating the Price of a Natural Monopoly

A firm is a natural monopoly because it has economies of scale (declining average and marginal costs) over its entire output range. If price were regulated to be P_c , the firm would lose money and go out of business. Setting the price at P_r yields the largest possible output consistent with the firm's remaining in business; excess profit is zero.

10.5 Monopsony

So far our discussion of market power has focused entirely on the seller side of the market. Now we turn to the *buyer* side. We will see that if there are not too many buyers, they can also have market power and use it profitably to affect the price they pay for a product.

First, a few terms.

- **Monopsony** refers to a market in which there is a single buyer.
- An **oligopsony** is a market with only a few buyers.
- With one or only a few buyers, some buyers may have **monopsony power**: a buyer's ability to affect the price of a good. Monopsony power enables the buyer to purchase the good for less than the price that would prevail in a competitive market.

Suppose you are trying to decide how much of a good to purchase. You could apply the basic marginal principle—keep purchasing units of the good until the last unit purchased gives additional value, or utility, just equal to the cost of that last unit. In other words, on the margin, additional benefit should just be offset by additional cost.

Let's look at this additional benefit and additional cost in more detail. We use the term **marginal value** to refer to the additional benefit from purchasing one more unit of a good. How do we determine marginal value? Recall from Chapter 4 that an individual demand curve determine marginal value, or marginal utility, as a function of the quantity purchased. Therefore, your *marginal value schedule* is your *demand* curve for the good. An individual's demand curve slopes downward because the marginal value obtained from buying one more unit of a good declines as the total quantity purchased increases.

The additional cost of buying one more unit of a good is called the **marginal expenditure**. What that marginal expenditure is depends on whether you are a competitive buyer or a buyer with monopsony power. Suppose you are a competitive buyer—in other words, you have no influence over the price of the good. In that case, the cost of each unit you buy is the same no matter how many units you purchase; it is the market price of the good. Figure 10.13(a) illustrates this principle. The price you pay per unit is your **average expenditure** per unit, and it is the same for all units. But what is your *marginal expenditure* per unit? As a competitive buyer, your marginal expenditure is equal to your average expenditure, which in turn is equal to the market price of the good.

Figure 10.13(a) also shows your marginal value schedule (i.e., your demand curve). How much of the good should you buy? You should buy until the marginal value of the last unit is just equal to the marginal expenditure on that unit. Thus you should purchase quantity Q^* at the intersection of the marginal expenditure and demand curves.

We introduced the concepts of marginal and average expenditure because they will make it easier to understand what happens when buyers have monopsony power. But before considering that situation, let's look at the analogy between competitive buyer conditions and competitive seller conditions. Figure 10.13(b) shows how a perfectly competitive seller decides how much to produce and sell. Because the seller takes the market price as given, both average and marginal revenue are equal to the price. The profit-maximizing quantity is at the intersection of the marginal revenue and marginal cost curves.

oligopsony Market with only a few buyers.

monopsony power Buyer's ability to affect the price of a good.

marginal value Additional benefit derived from purchasing one more unit of a good.

In §4.1, we explain that as we move down along a demand curve, the value the consumer places on an additional unit of the good falls.

marginal expenditure Additional cost of buying one more unit of a good.

average expenditure Price paid per unit of a good.

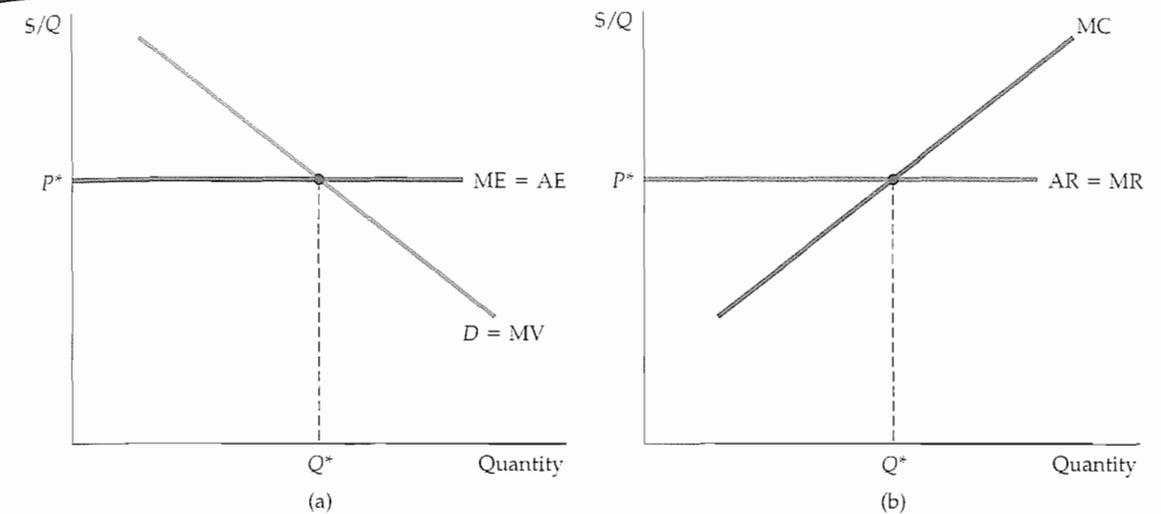


FIGURE 10.13 Competitive Buyer Compared to Competitive Seller

In (a), the competitive buyer takes market price P^* as given. Therefore, marginal expenditure and average expenditure are constant and equal; quantity purchased is found by equating price to marginal value (demand). In (b), the competitive seller also takes price as given. Marginal revenue and average revenue are constant and equal; quantity sold is found by equating price to marginal cost.

Now suppose that you are the *only* buyer of the good. Again you face a market supply curve, which tells you how much producers are willing to sell as a function of the price you pay. Should the quantity you purchase be at the point where your marginal value curve intersects the market supply curve? No. If you want to maximize your net benefit from purchasing the good, you should purchase a smaller quantity, which you will obtain at a lower price.

To determine how much to buy, set the marginal value from the last unit purchased equal to the marginal expenditure on that unit.¹⁴ Note, however, that the market supply curve is not the marginal expenditure curve. The market supply curve shows how much you must pay *per unit*, as a function of the total number of units you buy. In other words, the supply curve is the *average expenditure* curve. And because this average expenditure curve is upward sloping, the marginal expenditure curve must lie above it. The decision to buy an extra unit raises the price that must be paid for *all* units, not just the extra one.¹⁵

¹⁴Mathematically, we can write the net benefit NB from the purchase as $NB = V - E$, where V is the value to the buyer of the purchase and E is the expenditure. Net benefit is maximized when $\Delta NB/\Delta Q = 0$. Then

$$\Delta NB/\Delta Q = \Delta V/\Delta Q - \Delta E/\Delta Q = MV - ME = 0$$

so that $MV = ME$

¹⁵To obtain the marginal expenditure curve algebraically, write the supply curve with price on the left-hand side: $P = P(Q)$. Then total expenditure E is price times quantity, or $E = P(Q)Q$, and marginal expenditure is

$$ME = \Delta E/\Delta Q = P(Q) + Q(\Delta P/\Delta Q)$$

Because the supply curve is upward sloping, $\Delta P/\Delta Q$ is positive, and marginal expenditure is greater than average expenditure.

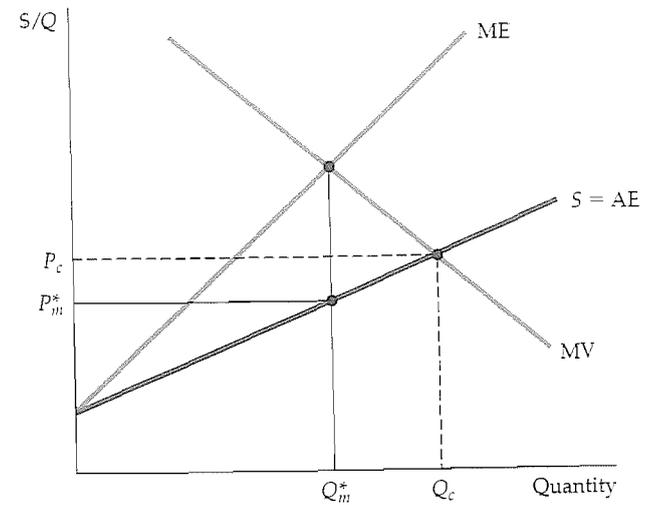


FIGURE 10.14 Monopsonist Buyer

The market supply curve is the monopsonist's average expenditure curve AE. Average expenditure is rising, so marginal expenditure lies above it. The monopsonist purchases quantity Q_m^* , where marginal expenditure and marginal value (demand) intersect. The price paid per unit P_m^* is then found from the average expenditure (supply) curve. In a competitive market, price and quantity, P_c and Q_c , are both higher. They are found at the point where average expenditure (supply) and marginal value (demand) intersect.

Figure 10.14 illustrates this principle. The optimal quantity for the monopsonist to buy, Q_m^* , is found at the intersection of the demand and marginal expenditure curves. The price that the monopsonist pays is found from the supply curve: It is the price P_m^* that brings forth the supply Q_m^* . Finally, note that this quantity Q_m^* is less, and the price P_m^* is lower, than the quantity and price that would prevail in a competitive market, Q_c and P_c .

Monopsony and Monopoly Compared

Monopsony is easier to understand if you compare it with monopoly. Figures 10.15(a) and 10.15(b) illustrate this comparison. Recall that a monopolist can charge a price above marginal cost because it faces a downward-sloping demand, or average revenue curve, so that marginal revenue is less than average revenue. Equating marginal cost with marginal revenue leads to a quantity Q^* that is less than what would be produced in a competitive market, and to a price P^* that is higher than the competitive price P_c .

The monopsony situation is exactly analogous. As Figure 10.15(b) illustrates, the monopsonist can purchase a good at a price below its marginal value because it faces an upward-sloping supply, or average expenditure, curve. Thus for a monopsonist, marginal expenditure is greater than average expenditure. Equating marginal value with marginal expenditure leads to a quantity Q^* that is less than what would be bought in a competitive market, and to a price P^* that is lower than the competitive price P_c .

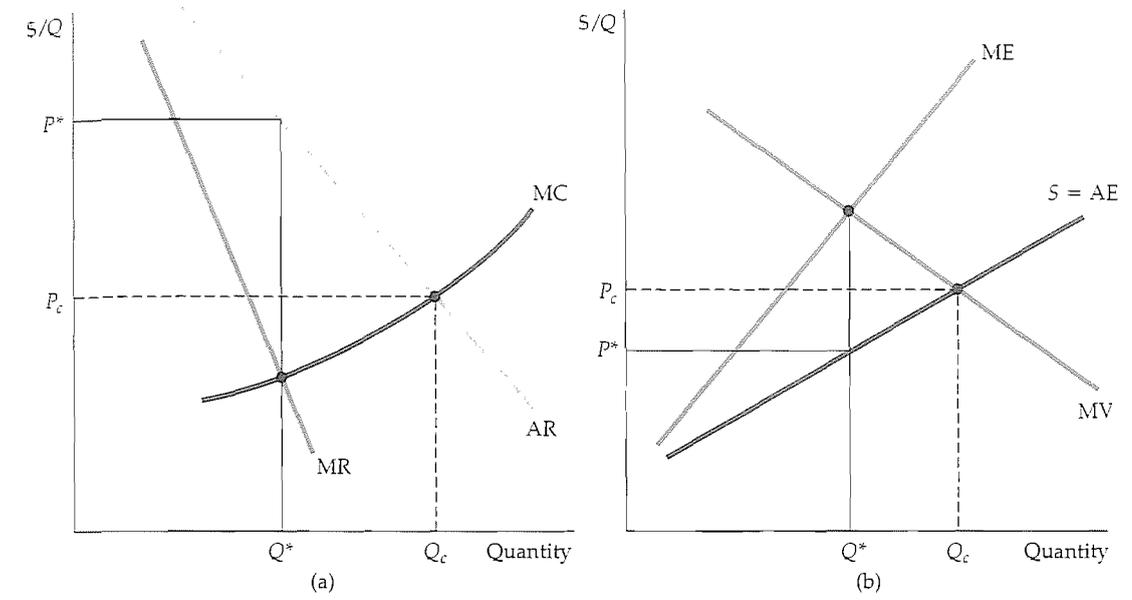


FIGURE 10.15 Monopoly and Monopsony

These diagrams show the close analogy between monopoly and monopsony. (a) The monopolist produces where marginal revenue intersects marginal cost. Average revenue exceeds marginal revenue, so that price exceeds marginal cost. (b) The monopsonist purchases up to the point where marginal expenditure intersects marginal value. Marginal expenditure exceeds average expenditure, so that marginal value exceeds price.

10.6 Monopsony Power

Much more common than pure monopsony are markets with only a few firms competing among themselves as buyers, so that each firm has some monopsony power. For example, the major U.S. automobile manufacturers compete with one another as buyers of tires. Because each of them accounts for a large share of the tire market, each has some monopsony power in that market. General Motors, the largest, might be able to exert considerable monopsony power when contracting for supplies of tires (and other automotive parts).

In a competitive market, price and marginal value are equal. A buyer with monopsony power, however, can purchase a good at a price below marginal value. The extent to which price is marked down below marginal value depends on the elasticity of supply facing the buyer.¹⁶ If supply is very elastic (E_s is large), the markdown will be small and the buyer will have little monopsony power. Conversely, if supply is very inelastic, the markdown will be large and the buyer will have considerable monopsony power. Figures 10.16(a) and 10.16(b) illustrate these two cases.

¹⁶ The exact relationship (analogous to equation (10.1)) is given by $(MV - P)/P = 1/E_s$. This equation follows because $MV = ME$ and $ME = \Delta(PQ)/\Delta Q = P + Q(\Delta P/\Delta Q)$.

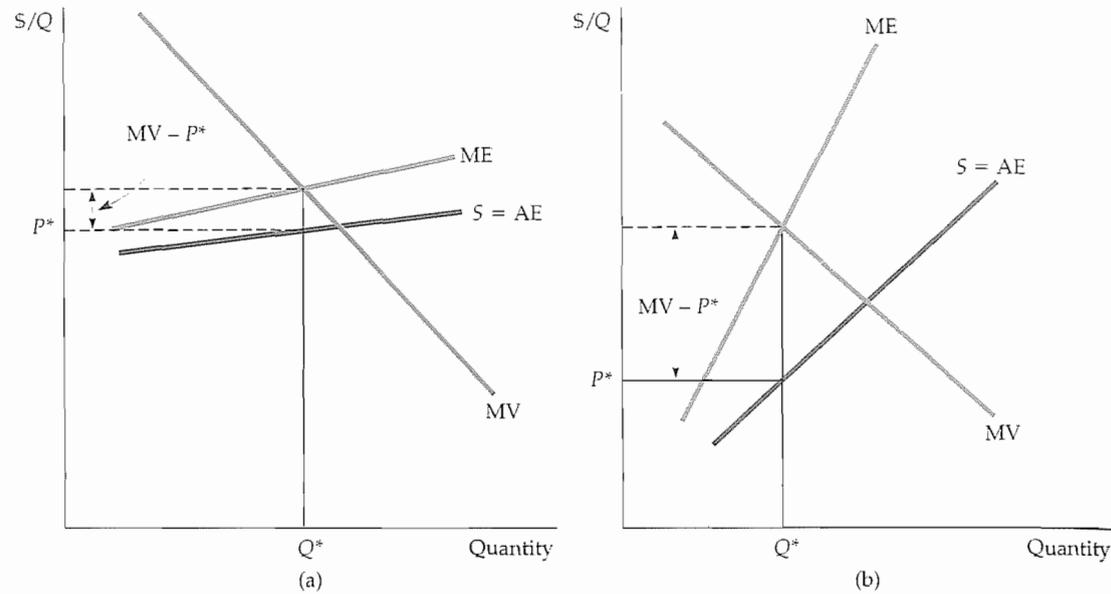


FIGURE 10.16 Monopsony Power: Elastic versus Inelastic Supply

Monopsony power depends on the elasticity of supply. When supply is elastic, as in (a), marginal expenditure and average expenditure do not differ by much, so price is close to what it would be in a competitive market. The opposite is true when supply is inelastic, as in (b).

Sources of Monopsony Power

What determines the degree of monopsony power in a market? Again, we can draw analogies with monopoly and monopoly power. We saw that monopoly power depends on three things: the elasticity of market demand, the number of sellers in the market, and how those sellers interact. Monopsony power depends on three similar things: The elasticity of market supply, the number of buyers in the market, and how those buyers interact.

Elasticity of Market Supply A monopsonist benefits because it faces an upward-sloping supply curve, so that marginal expenditure exceeds average expenditure. The less elastic the supply curve, the greater the difference between marginal expenditure and average expenditure and the more monopsony power the buyer enjoys. If only one buyer is in the market—a pure monopsonist—its monopsony power is completely determined by the elasticity of market supply. If supply is highly elastic, monopsony power is small and there is little gain in being the only buyer.

Number of Buyers Most markets have more than one buyer, and the number of buyers is an important determinant of monopsony power. When the number of buyers is very large, no single buyer can have much influence over price. Thus each buyer faces an extremely elastic supply curve, so that the market is almost completely competitive. The potential for monopsony power arises when the number of buyers is limited.

Interaction Among Buyers Finally, suppose three or four buyers are in the market. If those buyers compete aggressively, they will bid up the price close to

their marginal value of the product, and thus they will have little monopsony power. On the other hand, if those buyers compete less aggressively, or even collude, prices will not be bid up very much, and the buyers' degree of monopsony power might be nearly as high as if there were only one buyer.

So as with monopoly power, there is no simple way to predict how much monopsony power buyers will have in a market. We can count the number of buyers, and we can often estimate the elasticity of supply, but that is not enough. Monopsony power also depends on the interaction among buyers, which can be more difficult to ascertain.

The Social Costs of Monopsony Power

Because monopsony power results in lower prices and lower quantities purchased, we would expect it to make the buyer better off and sellers worse off. But suppose we value the welfare of buyers and sellers equally. How is aggregate welfare affected by monopsony power?

We can find out by comparing the consumer and producer surplus that results from a competitive market to the surplus that results when a monopsonist is the sole buyer. Figure 10.17 shows the average and marginal expenditure curves and marginal value curve for the monopsonist. The monopsonist's net benefit is maximized by purchasing a quantity Q_m at a price P_m such that marginal value equals marginal expenditure. In a competitive market, price equals marginal value. Thus the competitive price and quantity, P_c and Q_c , are found where the average expenditure and marginal value curves intersect. Now let's see how surplus changes if we move from the competitive price and quantity, P_c and Q_c , to the monopsony price and quantity, P_m and Q_m .

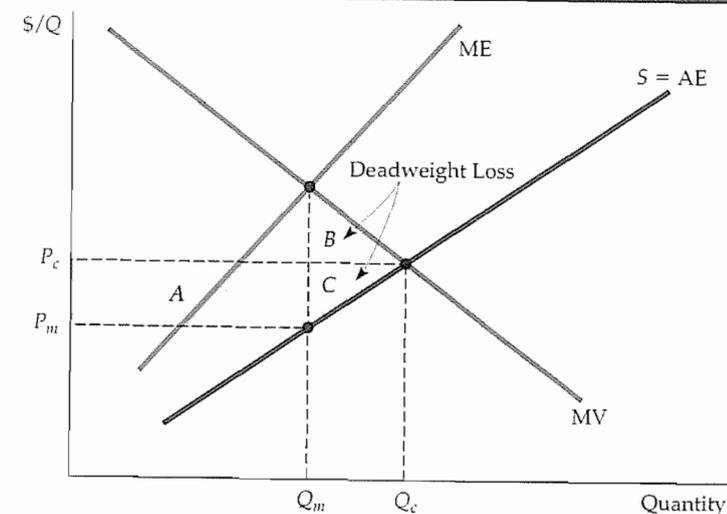


FIGURE 10.17 Deadweight Loss from Monopsony Power

The shaded rectangle and triangles show changes in consumer and producer surplus when moving from competitive price and quantity, P_c and Q_c , to monopsonist's price and quantity, P_m and Q_m . Because both price and quantity are lower, there is an increase in buyer (consumer) surplus given by $A - B$. Producer surplus falls by $A + C$, so there is a deadweight loss given by triangles B and C .

With monopsony, the price is lower and less is sold. Because of the lower price, sellers lose an amount of surplus given by rectangle *A*. In addition, sellers lose the surplus given by triangle *C* because of the reduced sales. The total loss of producer (seller) surplus is therefore $A + C$. The buyer gains the surplus given by rectangle *A* by buying at a lower price. However, the buyer buys less, Q_m instead of Q_c , and so loses the surplus given by triangle *B*. The total gain in surplus to the buyer is therefore $A - B$. Altogether, there is a net loss of surplus given by $B + C$. This is the *deadweight loss from monopsony power*. Even if the monopsonist's gains were taxed away and redistributed to the producers, there would be an inefficiency because output would be lower than under competition. The deadweight loss is the social cost of this inefficiency.

Bilateral Monopoly

bilateral monopoly Market with only one seller and one buyer.

What happens when a monopolist meets a monopsonist? It's hard to say. We call a market with only one seller and only one buyer a **bilateral monopoly**. If you think about such a market, you'll see why it is difficult to predict the price and quantity. Both the buyer and the seller are in a bargaining situation. Unfortunately, no simple rule determines which, if either, will get the better part of the bargain. One party might have more time and patience, or might be able to convince the other party that it will walk away if the price is too low or too high.

Bilateral monopoly is rare. Markets in which a few producers have some monopoly power and sell to a few buyers who have some monopsony power are more common. Although bargaining may still be involved, we can apply a rough principle here: *Monopsony power and monopoly power will tend to counteract each other*. In other words, the monopsony power of buyers will reduce the effective monopoly power of sellers, and vice versa. This does not mean that the market will end up looking perfectly competitive; if, for example, monopoly power is large and monopsony power small, the residual monopoly power would still be significant. But in general, monopsony power will push price closer to marginal cost, and monopoly power will push price closer to marginal value.

EXAMPLE 10.4 Monopsony Power in U.S. Manufacturing

Monopoly power, as measured by the price-cost margin $(P - MC)/P$, varies considerably across manufacturing industries in the United States. Some industries have price-cost margins close to zero, while in other industries the price-cost margins are as high as 0.4 or 0.5. These variations are due in part to differences in the determinants of monopoly power: In some industries market demand is more elastic than in others; some industries have more sellers than others; and in some industries, sellers compete more aggressively than in others. But something else can help explain these variations in monopoly power—differences in monopsony power among the firms' customers.

The role of monopsony power was investigated in a statistical study of 327 U.S. manufacturing industries.¹⁷ The study sought to determine the extent to which variations in price-cost margins could be attributed to variations in

¹⁷ The study was by Steven H. Lustgarten, "The Impact of Buyer Concentration in Manufacturing Industries," *Review of Economics and Statistics* 57 (May 1975): 125–32.

monopsony power by buyers in each industry. Although the degree of buyers' monopsony power could not be measured directly, data were available for variables that help determine monopsony power, such as buyer concentration (the fraction of total sales going to the three or four largest firms) and the average annual size of buyers' orders.

The study found that buyers' monopsony power had an important effect on the price-cost margins of sellers and could significantly reduce any monopoly power that sellers might otherwise have. Take, for example, the concentration of buyers, an important determinant of monopsony power. In industries where only four or five buyers account for all or nearly all sales, the price-cost margins of sellers would on average be as much as 10 percentage points lower than in comparable industries with hundreds of buyers accounting for sales.

A good example of monopsony power in manufacturing is the market for automobile parts and components, such as brakes and radiators. Each major car producer in the United States typically buys an individual part from at least three, and often as many as a dozen, suppliers. In addition, for a standardized product, such as brakes, each automobile company usually produces part of its needs itself, so that it is not totally reliant on outside firms. This puts companies like General Motors and Ford in an excellent bargaining position with respect to their suppliers. Each supplier must compete for sales against five or ten other suppliers, but each can sell to only a few buyers. For a specialized part, a single auto company may be the *only* buyer. As a result, the automobile companies have considerable monopsony power.

This monopsony power becomes evident from the conditions under which suppliers must operate. To obtain a sales contract, a supplier must have a track record of reliability, in terms of both product quality and ability to meet tight delivery schedules. Suppliers are also often required to respond to changes in volume, as auto sales and production levels fluctuate. Finally, pricing negotiations are notoriously difficult; a potential supplier will sometimes lose a contract because its bid is a penny per item higher than those of its competitors. Not surprisingly, producers of parts and components usually have little or no monopoly power.

10.7 Limiting Market Power: The Antitrust Laws

We have seen that market power—whether wielded by sellers or buyers—harms potential purchasers who could have bought at competitive prices. In addition, market power reduces output, which leads to a deadweight loss. Excessive market power also raises problems of equity and fairness: If a firm has significant monopoly power, it will profit at the expense of consumers. In theory, a firm's excess profits could be taxed away and redistributed to the buyers of its products, but such a redistribution is often impractical. It is difficult to determine what portion of a firm's profit is attributable to monopoly power, and it is even more difficult to locate all the buyers and reimburse them in proportion to their purchases.

How, then, can society limit market power and prevent it from being used anticompetitively? For a natural monopoly, such as an electric utility company, direct price regulation is the answer. But more generally, the answer is to prevent

antitrust laws Rules and regulations prohibiting actions that restrain, or are likely to restrain, competition.

firms from acquiring excessive market power in the first place, and to limit the use of that power if it is acquired. In the United States, this is done via the **antitrust laws**: a set of rules and regulations designed to promote a competitive economy by prohibiting actions that restrain, or are likely to restrain, competition, and by restricting the forms of market structure that are allowable.

Monopoly power can arise in a number of ways, each of which is covered by the antitrust laws. Section 1 of the Sherman Act (which was passed in 1890) prohibits contracts, combinations, or conspiracies in restraint of trade. One obvious example of an illegal combination is an explicit agreement among producers to restrict their outputs and/or “fix” price above the competitive level. There have been numerous instances of such illegal combinations. For example:

- In 1983, six companies and six executives were indicted for conspiring to fix the price of copper tubing over a six-year period.
- In 1996, Archer Daniels Midland Company (ADM) and two other major producers of lysine (an animal feed additive) pleaded guilty to criminal charges of price fixing. In 1999, three ADM executives were sentenced to prison terms ranging from two to three years for their roles in the price-fixing scheme.¹⁸
- In 1999, four of the world’s largest drug and chemical companies—Roche A.G. of Switzerland, BASF A.G. of Germany, Rhône-Poulenc of France, and Takeda Chemical Industries of Japan—were charged by the U.S. Department of Justice with taking part in a global conspiracy to fix the prices of vitamins sold in the United States. The companies pleaded guilty to price fixing and agreed to pay fines totaling more than \$1 billion.¹⁹

parallel conduct Form of implicit collusion in which one firm consistently follows actions of another.

Firm A and Firm B need not meet or talk on the telephone to violate Section 1 of the Sherman Act; *implicit* collusion in the form of **parallel conduct** can also be construed as violating the law. For example, if Firm B consistently follows Firm A’s pricing (parallel pricing), and if the firm’s conduct is contrary to what one would expect companies to do in the absence of collusion (such as raising prices in the face of decreased demand and over-supply), an implicit understanding may be inferred.²⁰

Section 2 of the Sherman Act makes it illegal to monopolize or to attempt to monopolize a market and prohibits conspiracies that result in monopolization. The Clayton Act (1914) did much to pinpoint the kinds of practices that are likely to be anticompetitive. For example, the Clayton Act makes it unlawful for a firm with a large market share to require the buyer or lessor of a good not to

¹⁸ In 1993, ADM and three other firms were charged with fixing carbon dioxide prices. In the lysine case, proof of the conspiracy came in part from tapes of meetings at which prices were set and market shares divided up. At one meeting with executives from Ajinomoto Company of Japan, another lysine producer, James Randall, then the president of ADM, said, “We have a saying at this company. Our competitors are our friends and our customers are our enemies.” See “Video Tapes Take Star Role at Archer Daniels Trial,” *New York Times*, August 4, 1998; “Three Sentenced in Archer Daniels Midland Case,” *New York Times*, July 10, 1999.

¹⁹ “Tearing Down The Facades of ‘Vitamins Inc.’,” *New York Times*, October 10, 1999.

²⁰ The Sherman Act applies to all firms that do business in the United States (to the extent that a conspiracy to restrain trade could affect U.S. markets). However, foreign governments (or firms operating under their government’s control) are not subject to the act, so OPEC need not fear the wrath of the Justice Department. Also, firms can collude with respect to *exports*. The Webb-Pomerene Act (1918) allows price fixing and related collusion with respect to export markets, as long as domestic markets are unaffected by such collusion. Firms operating in this manner must form a “Webb-Pomerene Association” and register it with the government.

buy from a competitor. It also makes it illegal to engage in **predatory pricing**—pricing designed to drive current competitors out of business and to discourage new entrants (so that the predatory firm can enjoy higher prices in the future).

Monopoly power can also be achieved by a merger of firms into a larger and more dominant firm, or by one firm acquiring or taking control of another firm by purchasing its stock. The Clayton Act prohibits mergers and acquisitions if they “substantially lessen competition” or “tend to create a monopoly.”

The antitrust laws also limit possible anticompetitive conduct by firms in other ways. For example, the Clayton Act, as amended by the Robinson-Patman Act (1936), makes it illegal to discriminate by charging buyers of essentially the same product different prices if those price differences are likely to injure competition. Even then, firms are not liable if they can show that the price differences were necessary to meet competition. (As we will see in the next chapter, price discrimination is a common practice. It becomes the target of antitrust action when buyers suffer economic damages and competition is reduced.)

Another important component of the antitrust laws is the *Federal Trade Commission Act* (1914, amended in 1938, 1973, 1975), which created the Federal Trade Commission (FTC). This act supplements the Sherman and Clayton acts by fostering competition through a whole set of prohibitions against unfair and anticompetitive practices, such as deceptive advertising and labeling, agreements with retailers to exclude competing brands, and so on. Because these prohibitions are interpreted and enforced in administrative proceedings before the FTC, the act provides broad powers that reach further than other antitrust laws.

The antitrust laws are actually phrased vaguely in terms of what is and what is not allowed. They are intended to provide a general statutory framework to give the Justice Department, the FTC, and the courts wide discretion in interpreting and applying them. This is important because it is difficult to know in advance what might be an impediment to competition. Such ambiguity creates a need for common law (i.e., the practice whereby courts interpret statutes) and supplemental provisions and rulings (e.g., by the FTC or the Justice Department).

Enforcement of the Antitrust Laws

The antitrust laws are enforced in three ways.

1. *Through the Antitrust Division of the Department of Justice.* As an arm of the executive branch, its enforcement policies closely reflect the view of the administration in power. As the result of an external complaint or an internal study, the department can institute a criminal proceeding, bring a civil suit, or both. The result of a criminal action can be fines for the corporation and fines or jail sentences for individuals. For example, individuals who conspire to fix prices or rig bids can be charged with a *felony* and, if found guilty, may be sentenced to jail—something to remember if you are planning to parlay your knowledge of microeconomics into a successful business career! Losing a civil action forces a corporation to cease its anticompetitive practices and often to pay damages.
2. *Through the administrative procedures of the Federal Trade Commission.* Again, action can result from an external complaint or from the FTC’s own initiative. Should the FTC decide that action is required, it can either request a voluntary understanding to comply with the law or seek a formal commission order requiring compliance.

predatory pricing Practice of pricing to drive current competitors out of business and to discourage new entrants in a market so that a firm can enjoy higher future profits.

3. *Through private proceedings.* Individuals or companies can sue for *treble (three-fold) damages* inflicted on their businesses or property. The possibility of having to pay treble damages can be a strong deterrent to would-be violators. Individuals or companies can also ask the courts for injunctions to force wrongdoers to cease anticompetitive actions.

U.S. antitrust laws are more stringent and far-reaching than those of most other countries. In fact, some people have argued that they have prevented American industry from competing effectively in international markets. The laws certainly constrain American business and may at times have put American firms at a disadvantage in world markets. But this must be weighed against their benefits: Antitrust laws have been crucial for maintaining competition, and competition is essential for economic efficiency, innovation, and growth.

EXAMPLE 10.5 A Phone Call About Prices

In 1981 and early 1982, American Airlines and Braniff Airways were competing fiercely with each other for passengers. A fare war broke out as the firms undercut each other's prices to capture market share. On February 21, 1982, Robert Crandall, president and CEO of American, made a phone call to Howard Putnam, president and chief executive of Braniff. To Crandall's later surprise, the call had been taped. It went like this:²¹

Crandall: I think it's dumb as hell for Christ's sake, all right, to sit here and pound the @!#\$%&! out of each other and neither one of us making a @!#\$%&! dime.

Putnam: Well . . .

Crandall: I mean, you know, @!#\$%&!, what the hell is the point of it?

Putnam: But if you're going to overlay every route of American's on top of every route that Braniff has—I just can't sit here and allow you to bury us without giving our best effort.

Crandall: Oh sure, but Eastern and Delta do the same thing in Atlanta and have for years.

Putnam: Do you have a suggestion for me?

Crandall: Yes, I have a suggestion for you. Raise your @!#\$%&! fares 20 percent. I'll raise mine the next morning.

Putnam: Robert, we . . .

Crandall: You'll make more money and I will, too.

Putnam: We can't talk about pricing!

Crandall: Oh @!#\$%&!, Howard. We can talk about any @!#\$%&! thing we want to talk about.

Crandall was wrong. Corporate executives cannot talk about anything they want. Talking about prices and agreeing to fix them is a clear violation of Section 1 of the Sherman Act. Putnam must have known this because he promptly rejected Crandall's suggestion. After learning about the call, the Justice Department filed a suit accusing Crandall of violating the antitrust laws by proposing to fix prices.

²¹ According to the *New York Times*, February 24, 1983.

However, *proposing* to fix prices is not enough to violate Section 1 of the Sherman Act: For the law to be violated, the two parties must *agree* to collude. Therefore, because Putnam had rejected Crandall's proposal, Section 1 was not violated. The court later ruled, however, that a proposal to fix prices could be an attempt to monopolize part of the airline industry and, if so, would violate Section 2 of the Sherman Act. American Airlines promised the Justice Department never again to engage in such activity.

EXAMPLE 10.6 The United States versus Microsoft

Over the past decade, Microsoft Corporation has grown to become the largest computer software company in the world. Its Windows operating system has over 90 percent of the worldwide market for personal computer operating systems. Microsoft also dominates the office productivity market: Its Office Suite, which includes Word (word processing), Excel (spreadsheets), and Powerpoint (presentations) held over a 90 percent worldwide market share in 1999.

Microsoft's incredible success has been due in good part to the creative technological and marketing decisions of the company and its CEO, Bill Gates. Is there anything wrong as a matter of either economics or law with being so successful and dominant? It all depends. Under the antitrust laws, efforts by firms to restrain trade or to engage in activities that inappropriately maintain monopolies are illegal. Did Microsoft engage in anticompetitive, illegal practices?

The U.S. Government says yes; Microsoft disagrees. In October 1998, the Antitrust Division of the U.S. Department of Justice (DOJ) put Microsoft's behavior to the test: It filed suit, raising a broad set of issues that created the most significant antitrust law suit of the past two decades. The ensuing trial ended in June 1999, but in the absence of any settlement between the government and Microsoft, the final chapter of the story is unlikely to be written for years to come. Here is a brief road map of some of the DOJ's major claims and Microsoft's response.

- **DOJ claim:** Microsoft has a great deal of market power in the market for PC operating systems—enough to meet the legal definition of monopoly power.
- **MS response:** Microsoft does not meet the legal test for monopoly power because it faces significant threats from potential competitors that offer or will offer platforms to compete with Windows.
- **DOJ claim:** Microsoft viewed Netscape's Internet browser (Netscape Navigator) as a threat to its monopoly over the PC operating system market. The threat exists because Netscape's browser includes Sun's Java software, which can run programs that have been written for *any* operating system, including those that compete with Windows, such as Apple, Unix, and Linux. In violation of Section 1 of the Sherman Act, Microsoft entered into exclusionary agreements with computer manufacturers, Internet service providers, and Internet content providers with the objective of raising the cost to Netscape of making its browser available to consumers. This action impaired Netscape's ability to compete fairly with Microsoft's Internet Explorer for the browser business.

- **MS response:** The contracts were not unduly restrictive. In any case, Microsoft unilaterally agreed to stop most of them.
- **DOJ claim:** In violation of Section 2 of the Sherman Act, Microsoft engaged in practices designed to maintain its monopoly in the market for desktop PC operating systems. Most importantly, it tied its browser to the Windows 98 operating system, even though doing so was technically unnecessary and provides little or no benefit to consumers. This action is predatory because it makes it difficult or impossible for Netscape and other firms to successfully offer competing products.
- **MS response:** There are benefits to incorporating the browser functionality into the operating system. Not being allowed to integrate new functionality into an operating system will discourage innovation. Offering consumers a choice between separate or integrated browsers would cause confusion in the marketplace.
- **DOJ claim:** In violation of Section 2 of the Sherman Act, Microsoft attempted to divide the browser business with Netscape and engaged in similar conduct with both Apple Computer and Intel.
- **MS response:** Microsoft's meetings with Netscape, Apple, and Intel were for valid business reasons. Indeed, it is useful for consumers and firms to agree on common standards and protocols in developing computer software.

These are only a few of the highlights of an eight-month trial that was hard fought on a range of economic topics covering a wide array of antitrust issues. To learn more about the case, look at the Web sites of the two parties: www.usdoj.gov/atd and www.microsoft.com.

SUMMARY

1. Market power is the ability of sellers or buyers to affect the price of a good.
2. Market power comes in two forms. When sellers charge a price that is above marginal cost, we say that they have monopoly power, which we measure by the extent to which price exceeds marginal cost. When buyers can obtain a price below their marginal value of the good, we say they have monopsony power, which we measure by the extent to which marginal value exceeds price.
3. Monopoly power is determined in part by the number of firms competing in the market. If there is only one firm—a pure monopoly—monopoly power depends entirely on the elasticity of market demand. The less elastic the demand, the more monopoly power the firm will have. When there are several firms, monopoly power also depends on how the firms interact. The more aggressively they compete, the less monopoly power each firm will have.
4. Monopsony power is determined in part by the number of buyers in the market. If there is only one buyer—a pure monopsony—monopsony power depends on the elasticity of market supply. The less elastic the supply, the more monopsony power the buyer will have. When there are several buyers, monopsony power also depends on how aggressively they compete for supplies.
5. Market power can impose costs on society. Because monopoly and monopsony power both cause production to fall below the competitive level, there is a deadweight loss of consumer and producer surplus. There can be additional social costs from rent seeking.
6. Sometimes, scale economies make pure monopoly desirable. But the government will still want to regulate price to maximize social welfare.
7. More generally, we rely on the antitrust laws to prevent firms from obtaining excessive market power.

QUESTIONS FOR REVIEW

1. A monopolist is producing at a point at which marginal cost exceeds marginal revenue. How should it adjust its output to increase profit?
2. We write the percentage markup of price over marginal cost as $(P - MC)/P$. For a profit-maximizing monopolist, how does this markup depend on the elasticity of demand? Why can this markup be viewed as a measure of monopoly power?
3. Why is there no market supply curve under conditions of monopoly?
4. Why might a firm have monopoly power even if it is not the only producer in the market?
5. What are some of the sources of monopoly power? Give an example of each.
6. What factors determine the amount of monopoly power an individual firm is likely to have? Explain each one briefly.
7. Why is there a social cost to monopoly power? If the gains to producers from monopoly power could be redistributed to consumers, would the social cost of monopoly power be eliminated? Explain briefly.
8. Why will a monopolist's output increase if the government forces it to lower its price? If the government wants to set a price ceiling that maximizes the monopolist's output, what price should it set?
9. How should a monopolist decide how much of a product to buy? Will it buy more or less than a competitive buyer? Explain briefly.
10. What is meant by the term "monopsony power"? Why might a firm have monopsony power even if it is not the only buyer in the market?
11. What are some sources of monopsony power? What determines the amount of monopsony power an individual firm is likely to have?
12. Why is there a social cost to monopsony power? If the gains to buyers from monopsony power could be redistributed to sellers, would the social cost of monopsony power be eliminated? Explain briefly.
13. How do the antitrust laws limit market power in the United States? Give examples of major provisions of the laws.
14. Explain briefly how the U.S. antitrust laws are actually enforced.

EXERCISES

1. Will an increase in the demand for a monopolist's product always result in a higher price? Explain. Will an increase in the supply facing a monopsonist buyer always result in a lower price? Explain.
2. Caterpillar Tractor, one of the largest producers of farm machinery in the world, has hired you to advise them on pricing policy. One of the things the company would like to know is how much a 5-percent increase in price is likely to reduce sales. What would you need to know to help the company with this problem? Explain why these facts are important.
3. A monopolist firm faces a demand with constant elasticity of -2.0 . It has a constant marginal cost of \$20 per unit and sets a price to maximize profit. If marginal cost should increase by 25 percent, would the price charged also rise by 25 percent?
4. A firm faces the following average revenue (demand) curve:
 - a. What is the level of production, price, and total profit per week?
 - b. If the government decides to levy a tax of 10 cents per unit on this product, what will be the new level of production, price, and profit?
5. The following table shows the demand curve facing a monopolist who produces at a constant marginal cost of \$10:

PRICE	QUANTITY
27	0
24	2
21	4
18	6
15	8
12	10
9	12
6	14
3	16
0	18

- $P = 100 - 0.01Q$
- where Q is weekly production and P is price, measured in cents per unit. The firm's cost function is given by $C = 50Q + 30,000$. Assume that the firm maximizes profits.
- a. Calculate the firm's marginal revenue curve.