

PREDICTIVE ANALYTICS FOR BUSINESS STRATEGY

Reasoning from Data to Actionable Knowledge



Mc
Graw
Hill
Education

JEFFREY T. PRINCE

PREDICTIVE ANALYTICS FOR BUSINESS STRATEGY

Reasoning from Data to Actionable Knowledge



Predictive Analytics for Business Strategy:

REASONING FROM DATA TO
ACTIONABLE KNOWLEDGE



THE MCGRAW-HILL SERIES ECONOMICS

ESSENTIALS OF ECONOMICS

Brue, McConnell, and Flynn

Essentials of Economics

Fourth Edition

Mandel

Economics: The Basics

Third Edition

Schiller

Essentials of Economics

Tenth Edition

PRINCIPLES OF ECONOMICS

Asarta and Butters

Principles of Economics,

Principles of Microeconomics,

Principles of Macroeconomics

Second Edition

Colander

Economics, Microeconomics, and Macroeconomics

Tenth Edition

Frank, Bernanke, Antonovics, and Heffetz

Principles of Economics,

Principles of Microeconomics,

Principles of Macroeconomics

Seventh Edition

Frank, Bernanke, Antonovics, and Heffetz

Streamlined Editions: Principles of Economics, Principles of Microeconomics,
Principles of Macroeconomics

Third Edition

Karlan and Morduch

Economics, Microeconomics, and Macroeconomics

Second Edition

McConnell, Brue, and Flynn

Economics, Microeconomics, Macroeconomics

Twenty-First Edition

McConnell, Brue, and Flynn

Brief Editions: Microeconomics and Macroeconomics

Second Edition

Samuelson and Nordhaus

Economics, Microeconomics, and Macroeconomics

Nineteenth Edition

Schiller

The Economy Today, The Micro Economy Today, and The Macro Economy Today

Fifteenth Edition

Slavin

Economics, Microeconomics, and Macroeconomics

Eleventh Edition

ECONOMICS OF SOCIAL ISSUES

Guell

Issues in Economics Today

Eighth Edition

Register and Grimes

Economics of Social Issues

Twenty-First Edition

ECONOMETRICS AND DATA ANALYSIS

Gujarati and Porter

Basic Econometrics

Fifth Edition

Gujarati and Porter

Essentials of Econometrics

Fourth Edition

Hilmer and Hilmer

Practical Econometrics

First Edition

Prince

Predictive Analytics for Business Strategy

First Edition

MANAGERIAL ECONOMICS

Baye and Prince

Managerial Economics and Business Strategy

Ninth Edition

Brickley, Smith, and Zimmerman

Managerial Economics and Organizational Architecture

Sixth Edition

Thomas and Maurice

Managerial Economics

Twelfth Edition

INTERMEDIATE ECONOMICS

Bernheim and Whinston

Microeconomics

Second Edition

Dornbusch, Fischer, and Startz

Macroeconomics

Twelfth Edition

Frank

Microeconomics and Behavior

Ninth Edition

ADVANCED ECONOMICS

Romer

Advanced Macroeconomics

Fourth Edition

MONEY AND BANKING

Cecchetti and Schoenholtz

Money, Banking, and Financial Markets

Fifth Edition

URBAN ECONOMICS

O'Sullivan

Urban Economics

Eighth Edition

LABOR ECONOMICS

Borjas

Labor Economics

Seventh Edition

McConnell, Brue, and Macpherson

Contemporary Labor Economics
Eleventh Edition

PUBLIC FINANCE

Rosen and Gayer

Public Finance
Tenth Edition

ENVIRONMENTAL ECONOMICS

Field and Field

Environmental Economics: An Introduction
Seventh Edition

INTERNATIONAL ECONOMICS

Appleyard and Field

International Economics
Ninth Edition

Pugel

International Economics
Sixteenth Edition

Predictive Analytics for Business Strategy:

**REASONING FROM DATA TO
ACTIONABLE KNOWLEDGE**

Jeffrey T. Prince

*Professor of Business Economics & Public Policy
Harold A. Poling Chair in Strategic Management
Kelley School of Business
Indiana University*





PREDICTIVE ANALYTICS FOR BUSINESS STRATEGY: REASONING FROM DATA TO ACTIONABLE KNOWLEDGE

Published by McGraw-Hill Education, 2 Penn Plaza, New York, NY 10121. Copyright © 2019 by McGraw-Hill Education. All rights reserved. Printed in the United States of America. No part of this publication may be reproduced or distributed in any form or by any means, or stored in a database or retrieval system, without the prior written consent of McGraw-Hill Education, including, but not limited to, in any network or other electronic storage or transmission, or broadcast for distance learning.

Some ancillaries, including electronic and print components, may not be available to customers outside the United States.

This book is printed on acid-free paper.

1 2 3 4 5 6 7 8 9 LCR 21 20 19 18

ISBN 978-1-259-19151-0

MHID 1-259-19151-6

Executive Portfolio Manager: *Kathleen Hoenicke*

Senior Product Developer: *Christina Kouvelis*

Marketing Manager: *Bobby Pearson*

Content Project Managers: *Harvey Yep (Core) / Bruce Gin (Assessment)*

Senior Buyer: *Sandy Ludovissy*

Senior Designer: *Matt Diamond*

Senior Content Licensing Specialists: *Beth Thole (Text and Image)*

Cover Image: ©naqiewei/Getty Images

Compositor: *MPS Limited*

All credits appearing on page or at the end of the book are considered to be an extension of the copyright page.

Library of Congress Cataloging-in-Publication Data

Cataloging-in-Publication Data has been requested from the Library of Congress

The Internet addresses listed in the text were accurate at the time of publication. The inclusion of a website does not indicate an endorsement by the authors or McGraw-Hill Education, and McGraw-Hill Education does not guarantee the accuracy of the information presented at these sites.

mheducation.com/highered

To Mom and Dad
—Jeffrey T. Prince

about the author



Jeffrey T. Prince is Professor of Business Economics & Public Policy and Harold A. Poling Chair in Strategic Management at Indiana University's Kelley School of Business. He received his BA, in economics and BS, in mathematics and statistics from Miami University in 1998 and earned a PhD in economics from Northwestern University in 2004. Prior to joining Indiana University, he taught graduate and undergraduate courses at Cornell University.

Jeff has won top teaching honors as a faculty member at both Indiana University and Cornell, and as a graduate student at Northwestern. He has a broad research agenda within applied economics, having written and published on topics that include demand in technology and telecommunications markets, Internet diffusion, regulation in health care, risk aversion in insurance markets, and quality competition among airlines. He is one of a small number of economists to have published in both the top journal in economics (*American Economic Review*) and the top journal in management (*Academy of Management Journal*). Professor Prince currently is a co-editor at the *Journal of Economics and Management Strategy*, and serves on the editorial board for *Information Economics and Policy*.

In his free time, Jeff enjoys activities ranging from poker and bridge to running and racquetball.

preface

This book is meant to teach students how data analysis can inform strategy, within a framework centered on logical reasoning and practical communication.

The inspiration for this project comes from having taught for more than 20 years at the college level to a wide range of students (in mathematics and economics departments, and in business schools), covering a wide range of quantitative topics. During that time, it has become clear to me that the average business student recognizes, in principle, that quantitative skills are valuable. However, in practice (s)he often finds those skills intimidating and esoteric, wondering how exactly they will be useful in the workforce.

As of this writing, there are many econometrics books and many operations/data mining/business analytics books in the market. However, these books are generally geared toward the specialist, who needs to know the full methodological details. Hence, they are not especially approachable or appealing to the business student looking for a conceptual, broad-based understanding of the material. And in their design, it can be easy for students—specialists and nonspecialists alike—to “lose sight of the forest for the trees.”

As I see it, the problem with regard to data analysis is as follows: There is a large group of future businesspeople, both future analysts and managers, who recognize data analysis can be valuable. However, taking a course that is essentially a treatise on methodology and statistics causes the future managers to narrow their view toward simply “getting through” the course. In contrast, the future analysts may enjoy the material and often emerge understanding the methods and statistics, but lacking key skills to communicate and explain to managers what their results *mean*.

This book is designed to address the problem of the dual audience, by focusing on the role of data analysis in forming business strategy via predictive analytics. I chose this focus since all businesses, and virtually all management-level employees, must be mindful of the strategies they are following. Assessing the relative merit among a set of potential strategic moves generally requires one to forecast their future implications, often using data. Further, this component of predictive analytics contributes toward development of critical thinking about analytical findings. Both inside and outside of business, we are bombarded with statements with the following flavor: “If you do X, you should expect Y to happen.” (Commercials about the impact of switching insurance providers immediately come to mind.) A deep understanding of how data can inform strategy through predictive analytics will allow students to critically assess such statements.

Given its purpose, I believe this book can be the foundation of a course that will benefit both future analysts and managers. The course will give managers a basic understanding of what data can do in an important area of business (strategy formation) and present it in a way that doesn’t feel like a taxonomy of models and their statistical properties. Managers will thus develop a deeper understanding of the fundamental reasoning behind how and why data analysis can generate actionable knowledge, and be able to think critically about whether a given analysis has merit or not. Consequently, this course could provide future managers some valuable data training without forcing them to take a highly technical econometrics or data mining course. It will also serve as a natural complement to the strategy courses they take.

This course will give future analysts a bigger-picture understanding of what their analysis is trying to accomplish, and the conditions under which it can be deemed successful. It will also give them tools to better reason through these ideas and communicate them to others. Hence, it will serve as a valuable complement to the other, more technically focused, analytics courses they take.

KEY PEDAGOGICAL FEATURES

This text includes many features designed to ease the learning experience for students and the teaching process for instructors.

Data Challenges Each chapter opens by presenting a challenging data situation. In order for students to properly and effectively rise to the challenge, they must understand the material presented in that chapter.

At the end of the chapter, a concluding section titled *Rising to the Data Challenge* discusses how the challenge can be confronted and overcome using some of the newly acquired knowledge and skills that chapter develops.

These challenges, which bookend the chapter, are designed to motivate the reader to acquire the necessary skills by learning the chapter's material and understanding how to apply it.

Learning Objectives Learning objectives in each chapter organize the chapter content and enhance the learning experience.

Communicating Data Through real-world applications or explanations of text material in "layman's terms," *Communicating Data* examples demonstrate how to describe and explain data, data methods, and/or data results in a clear, intuitive manner. These

viii

examples are designed to enhance the reader's ability to communicate with a wide audience about data issues.

Reasoning Boxes *Reasoning Boxes* summarize main concepts from the text in the context of deductive and inductive reasoning. By understanding reasoning structure, readers will be better equipped to draw and explain their own conclusions using data and to properly critique others' data-based conclusions.

Demonstration Problems Beyond the opening *Data Challenge*, each chapter includes *Demonstration Problems* that help target and develop particular data skills. These are largely focused on primary applications of chapter material.

Key Terms and Marginal Definitions Each chapter ends with a list of key terms and concepts. These provide an easy way for instructors to assemble material covered in each chapter and for students to check their mastery of terminology. In addition, marginal definitions will appear as signposts throughout the text.

End-of-Chapter Problems Each chapter ends with two types of problems to test students' mastery of the material. First are *Conceptual Questions*, which test students' conceptual understanding of the material and demand pertinent communication and reasoning skills. Second are *Quantitative Problems*, which test students' ability to execute and explain (within a logical framework) pertinent data analytical methods. The Quantitative Problems are supported by Excel datasets available through McGraw-Hill Connect®.

Applications The material in this book is sufficient for any course that exclusively uses quizzes, homework, and/or exams for evaluation. However, to allow for a more enhanced, and applied, understanding of the material, the book concludes with an Applications section. This section has three parts. The first is "Critical Analysis of Data-Driven Conclusions." This section presents several real-world data applications that explicitly or implicitly lead to actionable conclusions, and then challenges students to critically assess these conclusions using the reasoning and data knowledge presented throughout the book. The second section is "Written Explanations of Data Analysis and Active Predictions." This section presents students with several mini-cases of data output, and challenges them to examine and explain the output in writing with appropriate reasoning. The third section is "Projects: Combining Analysis with Reason-based Communication." This last section provides three versions of a mini-project, based on projects Professor Prince has assigned in his own classes for several years. These projects require students to work from dataset to conclusions in a controlled, but realistic, environment. The projects are accompanied by datasets in Excel format, which can be easily tailored to instructors' needs. A key merit of these projects is flexibility, in that they can be used for individual- and/or group-level assessment.

ORGANIZED LEARNING

CHAPTER LEARNING OBJECTIVES

The organization of each chapter reflects common themes outlined by six to eight learning objectives listed at the beginning of each chapter. These objectives, along with AACSB and Bloom's taxonomy learning categories, are connected to the end-

of-chapter material and test bank questions to offer a comprehensive and thorough teaching and learning experience.

ASSURANCE OF LEARNING READY

Many educational institutions today are focused on the notion of *assurance of learning*, an important element of some accreditation standards. *Predictive Analytics for Business Strategy* is designed specifically to support your assurance of learning initiatives with a simple, yet powerful solution.

Instructors can use *Connect* to easily query for learning outcomes/objectives that directly relate to the learning objectives of the course. You can then use the reporting features of *Connect* to aggregate student results in similar fashion, making the collection and presentation of assurance of learning data simple and easy.

AACSB STATEMENT

McGraw-Hill Global Education is a proud corporate member of AACSB International. Understanding the importance and value of AACSB accreditation, *Predictive Analytics for Business Strategy* has sought to recognize the curricula guidelines detailed in the AACSB standards for business accreditation by connecting questions in the test bank and end-of-chapter material to the general knowledge and skill guidelines found in the AACSB standards.

It is important to note that the statements contained in *Predictive Analytics for Business Strategy* are provided only as a guide for the users of this text. The AACSB leaves content coverage and assessment within the purview of individual schools, the mission of the school, and the faculty. While *Predictive Analytics for Business Strategy* and the teaching package make no claim of any specific AACSB qualification or evaluation, we have labeled questions according to the general knowledge and skill areas.

MCGRAW-HILL CUSTOMER CARE CONTACT INFORMATION

At McGraw-Hill, we understand that getting the most from new technology can be challenging. That's why our services don't stop after you purchase our products.

You can e-mail our Product Specialists 24 hours a day to get product training online. Or you can search our knowledge bank of Frequently Asked Questions on our support website. For Customer Support, call **800-331-5094**, or visit www.mhhe.com/support. One of our Technical Support Analysts will be able to assist you in a timely fashion.

acknowledgments

I would like to thank the following reviewers, as well as hundreds of students at Indiana University's Kelley School of Business and colleagues who unselfishly gave up their own time to provide comments and suggestions to improve this book.

Imam Alam

University of Northern Iowa

Ahmad Bajwa

University of Arkansas at Little Rock

Steven Bednar

Elon University

Hooshang M. Beheshti

Radford University

Anton Bekkerman

Montana State University

Khurram S. Bhutta

Ohio University

Gary Black

University of Southern Indiana

Andre Boik

University of California, Davis

Ambarish Chandra

University of Toronto

Richard Cox

Arizona State University

Steven Cuellar

Sonoma State University

Craig Depken

University of North Carolina, Charlotte

Mark Dobeck

Cleveland State University

Tim Dorr

University of Bridgeport

Neal Duffy

State University of New York at Plattsburgh

Jerry Dunn

Southwestern Oklahoma State University

Kathryn Ernstberger

Indiana University Southeast

Ana L. Rosado Feger

Ohio University

Frederick Floss

Buffalo State University

Chris Forman

Georgia Institute of Technology

Avi Goldfarb

University of Toronto

Michael Gordinier
Washington University, St. Louis

Gauri Guha
Arkansas State University

Kuang-Chung Hsu
University of Central Oklahoma

Kyle Huff
Georgia Gwinnett College

Jongsung Kim
Bryant University

Ching-Chung Kuo
University of North Texas

Lirong Liu
Texas A&M University, Commerce

Stanislav Manonov
Montclair State University

John Mansuy
Wheeling Jesuit University

Ryan McDevitt
Duke University

Alex Meisami
Indiana University South Bend

Ignacio Molina
Arizona State University

Georgette Nicolaides
Syracuse University

Jie Peng

St. Ambrose University

Jeremy Petranka

Duke University

Kamelia Petrova

State University of New York at Plattsburgh

Claudia Pragman

Minnesota State University, Mankato

Reza Ramazani

Saint Michael's College

Doug Redington

Elon University

Sunil Sapra

California State University, Los Angeles

Robert Seamans

New York University

Mary Ann Shifflet

University of Southern Indiana

Timothy Simcoe

Boston University

Shweta Singh

Kean University

John Louis Sparco

Wilmington University

Arun Srinivasan

Indiana University Southeast

Purnima Srinivasan

Kean University

Leonie Stone

State University of New York at Geneseo

Richard Szal

Northern Arizona University

Vicar Valencia

Indiana University South Bend

Timothy S. Vaughan

University of Wisconsin, Eau Claire

Bindiganavale Vijayaraman

The University of Akron

Padmal Vitharana

Syracuse University

Razvan Vlaicu

University of Maryland

Rubina Vohra

New Jersey City University

Emily Wang

University of Massachusetts, Amherst

Miao Wang

Marquette University

Matthew Weinberg

Drexel University

Andy Welki

John Carroll University

John Whitehead

Appalachian State University

Peter Wui

University of Arkansas, Pine Bluff



McGraw-Hill Connect® is a highly reliable, easy-to-use homework and learning management solution that utilizes learning science and award-winning adaptive tools to improve student results.

Homework and Adaptive Learning

- Connect's assignments help students contextualize what they've learned through application, so they can better understand the material and think critically.
- Connect will create a personalized study path customized to individual student needs through SmartBook®.
- SmartBook helps students study more efficiently by delivering an interactive reading experience through adaptive highlighting and review.

Over 7 billion questions have been answered, making McGraw-Hill Education products more intelligent, reliable, and precise.

Connect's Impact on Retention Rates, Pass Rates, and Average Exam Scores



Using Connect improves retention rates by 19.8%, passing rates by 12.7%, and exam scores by 9.1%.

Quality Content and Learning Resources

- Connect content is authored by the world's best subject matter experts, and is available to your class through a simple and intuitive interface.
- The Connect eBook makes it easy for students to access their reading material on smartphones and tablets. They can study on the go and don't need internet access to use the eBook as a reference, with full functionality.
- Multimedia content such as videos, simulations, and games drive student engagement and critical thinking skills.

73% of instructors who use Connect require it; instructor satisfaction increases by 28% when Connect is required.



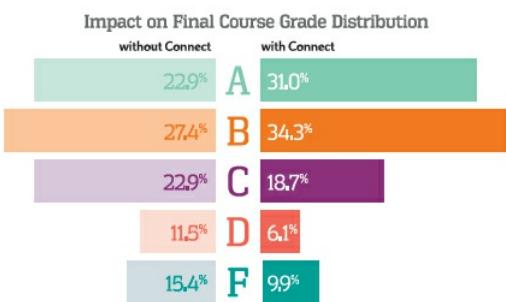
©McGraw-Hill Education

Robust Analytics and Reporting

- Connect Insight® generates easy-to-read reports on individual students, the class as a whole, and on specific assignments.
- The Connect Insight dashboard delivers data on performance, study behavior, and effort. Instructors can quickly identify students who struggle and focus on material that the class has yet to master.
- Connect automatically grades assignments and quizzes, providing easy-to-read reports on individual and class performance.



©Hero Images/Getty Images



More students earn
As and Bs when they
use Connect.

Trusted Service and Support

- Connect integrates with your LMS to provide single sign-on and automatic syncing of grades. Integration with Blackboard®, D2L®, and Canvas also provides automatic syncing of the course calendar and assignment-level linking.
- Connect offers comprehensive service, support, and training throughout every phase of your implementation.
- If you're looking for some guidance on how to use Connect, or want to learn tips and tricks from super users, you can find tutorials as you work. Our Digital Faculty Consultants and Student Ambassadors offer insight into how to achieve the results you want with Connect.

www.mheducation.com/connect

Supplements

I understand that the reliability and accuracy of the book and the accompanying supplements are of the utmost importance. To that end, I have been personally involved in crafting and accuracy checking each of the supplements. The following ancillaries are available for quick download and convenient access via the instructor resource material available through *Connect*.

POWERPOINT PRESENTATION

Presentation slides incorporate both the fundamental concepts of each chapter and the graphs and figures essential to each topic. These slides can be edited, printed, or rearranged to fit the needs of your course.

SOLUTIONS MANUAL

This manual contains solutions to the end-of-chapter conceptual questions and quantitative problems.

TEST BANK

A comprehensive test bank offers hundreds of questions categorized by learning objective, AACSB learning category, Bloom's taxonomy objectives, and level of difficulty.

COMPUTERIZED TEST BANK

TestGen is a complete, state-of-the-art test generator and editing application software that allows instructors to quickly and easily select test items from McGraw Hill's test bank content. The instructors can then organize, edit and customize

questions and answers to rapidly generate tests for paper or online administration. Questions can include stylized text, symbols, graphics, and equations that are inserted directly into questions using built-in mathematical templates. TestGen's random generator provides the option to display different text or calculated number values each time questions are used. With both quick and simple test creation and flexible and robust editing tools, TestGen is a complete test generator system for today's educators.

ONLINE RESOURCES

Student supplements for *Predictive Analytics for Business Strategy* are available online at www.mhhe.com/prince1e. These include datasets for all Quantitative Problems and sample datasets for the Course Project.

brief contents

- | | |
|-------------------|--|
| chapter 1 | The Roles of Data and Predictive Analytics in Business |
| chapter 2 | Reasoning with Data |
| chapter 3 | Reasoning from Sample to Population |
| chapter 4 | The Scientific Method: The Gold Standard for Establishing Causality |
| chapter 5 | Linear Regression as a Fundamental Descriptive Tool |
| chapter 6 | Correlation vs. Causality in Regression Analysis |
| chapter 7 | Basic Methods for Establishing Causal Inference |
| chapter 8 | Advanced Methods for Establishing Causal Inference |
| chapter 9 | Prediction for a Dichotomous Variable |
| chapter 10 | Identification and Data Assessment |

APPLICATIONS *Data Analysis Critiques, Write-ups, and*

Projects

GLOSSARY



contents

CHAPTER 1

The Roles of Data and Predictive Analytics in Business

Data Challenge: Navigating a Data Dump

Introduction

Defining Data and Data Uses in Business

Data

Predictive Analytics within Business Analytics

Business Strategy

Predictive Analytics for Business Strategy

Data Features

Structured vs. Unstructured Data

The Unit of Observation

Data-generating Process

Basic Uses of Data Analysis for Business

Queries

Pattern Discovery

Causal Inference

Data Analysis for the Past, Present, and Future

Lag and Lead Information

Predictive Analytics

Active Prediction for Business Strategy Formation

Rising to the Data Challenge

Summary / Key Terms and Concepts / Conceptual Questions / Quantitative Problems

- **Communicating Data 1.1:** Is/Are Data Singular or Plural?
- **Communicating Data 1.2:** Elaborating on Data Types
- **Communicating Data 1.3:** Situational Batting Averages
- **Communicating Data 1.4:** Indirect Causal Relationships in Purse Knockoffs
- **Communicating Data 1.5:** Passive and Active Prediction in Politics and Retail

CHAPTER 2

Reasoning with Data

Data Challenge: Testing for Sex Imbalance

Introduction

What is Reasoning?

Deductive Reasoning

Definition and Examples

Empirically Testable Conclusions

Inductive Reasoning

Definition and Examples

Evaluating Assumptions

Selection Bias

Rising to the Data Challenge

Summary / Key Terms and Concepts / Conceptual Questions / Quantitative Problems

- Communicating Data 2.1: Deducing Guilt and Innocence
- Communicating Data 2.2: Inductive Reasoning via Customer Testimonies
- Communicating Data 2.3: Selection Bias in News Network Polls
- Reasoning Box 2.1: Direct Proof and Transposition
- Reasoning Box 2.2: Inductive Reasoning for Evaluating Assumptions
- Reasoning Box 2.3: Selection Bias in Inductive Reasoning

CHAPTER 3

Reasoning from Sample to Population

Data Challenge: Knowing All Your Customers by Observing a Few

Introduction

Distributions and Sample Statistics

xv

Distributions of Random Variables

Data Samples and Sample Statistics

The Interplay Between Deductive and Inductive Reasoning in Active Predictions

Rising to the Data Challenge

Summary / Key Terms and Concepts / Conceptual Questions / Quantitative Problems

- Communicating Data 3.1: What Can Political Polls Tell Us about the General Population?
- Communicating Data 3.2: Does Working at Work Make a Difference?

- Reasoning Box 3.1: The Distribution of the Sample Mean
- Reasoning Box 3.2: Confidence Intervals
- Reasoning Box 3.3: The Distribution of the Sample Mean for Hypothesized Population Mean
- Reasoning Box 3.4: Hypothesis Testing
- Reasoning Box 3.5: Reasoning in Active Predictions

CHAPTER 4

The Scientific Method: The Gold Standard for Establishing Causality

Data Challenge: Does Dancing Yield Dollars?

Introduction

The Scientific Method

Definition and Details

The Scientific Method and Causal Inference

Data Analysis Using the Scientific Method

Experimental Data vs. Non-Experimental Data

Examples of Nonexperimental Data in Business

Consequences of Using Nonexperimental Data to Estimate Treatment Effects

Rising to the Data Challenge

Summary / Key Terms and Concepts / Conceptual Questions / Quantitative Problems

- Communicating Data 4.1: Penicillin and the Scientific Method
- Communicating Data 4.2: The Effect of Banner Ad Features
- Communicating Data 4.3: Music Training and Intelligence
- Communicating Data 4.4: Marshmallows and Reliability

- Communicating Data 4.5: The Rewards of Rudeness
- Reasoning Box 4.1: The Treatment Effect
- Reasoning Box 4.2: The Distribution of Experimental Outcomes
- Reasoning Box 4.3: Hypothesis Test for the Treatment Effect
- Reasoning Box 4.4: Confidence Interval for the Treatment Effect

CHAPTER 5

Linear Regression as a Fundamental Descriptive Tool

Data Challenge: Where to Park Your Truck?

Introduction

The Regression Line for a Dichotomous Treatment

An Intuitive Approach

A Formal Approach

The Regression Line for a Multi-Level Treatment

An Intuitive Approach

A Formal Approach

Sample Moments and Least Squares

Regression for Multiple Treatments

Single vs. Multiple Treatments

Multiple Regression

What Makes Regression Linear?

Rising to the Data Challenge

Summary / Key Terms and Concepts / Conceptual Questions / Quantitative Problems

- Communicating Data 5.1: Regression Line Origins
- Communicating Data 5.2: Least Squares vs. Least Absolute Deviations
- Communicating Data 5.3: Regression for Ratings
- Reasoning Box 5.1: The Regression Line for a Dichotomous Treatment
- Reasoning Box 5.2: The Simple Regression Line
- Reasoning Box 5.3: Multiple Regression

xvi

CHAPTER 6

Correlation vs. Causality in Regression Analysis

Data Challenge: Where to Park Your Truck—Redux

Introduction

The Difference Between Correlation and Causality

Regression Analysis for Correlation

Regression and Sample Correlation

Regression and Population Correlation

Passive Prediction Using Regression

Regression Analysis for Causality

Regression and Causation

Linking Causal Regression to the Experimental Ideal

Active Prediction Using Regression

The Relevance of Model Fit for Passive and Active Prediction

Rising to the Data Challenge

Summary / Key Terms and Concepts / Conceptual Questions / Quantitative Problems

- **Communicating Data 6.1:** Physical Fitness and Academic Success
- **Communicating Data 6.2:** Experiments vs. Causal Regression Analysis
- **Communicating Data 6.3:** Will Drinking Fatty Milk Make You Fat?
- **Reasoning Box 6.1:** Consistency of Regression Estimators for Population Correlations
- **Reasoning Box 6.2:** Confidence Intervals for Correlational Regression Analysis
- **Reasoning Box 6.3:** Hypothesis Testing for Correlational Regression Analysis
- **Reasoning Box 6.4:** Equivalence of Population Regression Equation and Determining Function
- **Reasoning Box 6.5:** Consistency of Regression Estimators for Determining Functions
- **Reasoning Box 6.6:** Confidence Intervals for Parameters of a Determining Function
- **Reasoning Box 6.7:** Hypothesis Testing for Parameters of a Determining Function

CHAPTER 7

Basic Methods for Establishing Causal Inference

Data Challenge: Does Working Out at Work Make for a Happy Worker?

Introduction

Assessing Key Assumptions Within a Causal Model

Random Sample

No Correlation between Errors and Treatments

Control Variables

Definition and Illustration

Dummy Variables

Selecting Controls

Proxy Variables

Form of the Determining Function

Rising to the Data Challenge

Summary / Key Terms and Concepts / Conceptual Questions / Quantitative Problems

- Communicating Data 7.1: Is Education Going Up in Smoke?
- Communicating Data 7.2: Does GDP Growth Proxy Economic Climate?
- Communicating Data 7.3: Trouble with the (Laffer) Curve
- Reasoning Box 7.1: Criteria for a Good Control
- Reasoning Box 7.2: Criteria for a Good Proxy Variable
- Reasoning Box 7.3: Why Polynomials Do the Trick—the Weierstrass Theorem

CHAPTER 8

Advanced Methods for Establishing Causal Inference

Data Challenge: Do TV Ads Generate Web Traffic?

Introduction

Instrumental Variables

Definition and Illustration

Two-Stage Least Squares Regression

Evaluating Instruments

Exogeneity

Classic Applications of Instrumental Variables for Business

Panel Data Methods

Difference-in-Differences

The Fixed-Effects Model

Practical Applications of Panel Data Methods for Business

Rising to the Data Challenge

Summary / Key Terms and Concepts / Conceptual Questions / Quantitative Problems

- Communicating Data 8.1: Measuring the Impact of Broadband Expansion
- Communicating Data 8.2: Using Diff-in-Diff to Assess the Minimum Wage
- Communicating Data 8.3: Does Multimarket Contact Affect Airlines' On-Time Performance?
- Reasoning Box 8.1: Using an Instrumental Variable to Achieve Causal Inference via 2SLS
- Reasoning Box 8.2: When Does Diff-in-diff Regression Solve an Endogeneity Problem?
- Reasoning Box 8.3: Implications of the Fixed Effects Model

CHAPTER 9

Prediction for a Dichotomous Variable

Data Challenge: Changing the Offer to Change Your Odds

Introduction

Limited Dependent Variables

The Linear Probability Model

Definition and Interpretation

Merits and Shortcomings

Probit And Logit Models

Latent Variable Formulation

Marginal Effects

Estimation and Interpretation

Merits and Shortcomings

Rising to the Data Challenge

Summary / Key Terms and Concepts / Conceptual Questions / Quantitative Problems

- **Communicating Data 9.1:** How to Model and Predict Cord Cutting
- **Communicating Data 9.2:** Characterizing Endogeneity within a Linear Probability Model
- **Communicating Data 9.3:** The “Right” Model for a Dichotomous Dependent Variable
- **Reasoning Box 9.1:** Interpretation of a Linear Probability Model
- **Reasoning Box 9.2:** Contrasting the Probit and Logit Model
- **Reasoning Box 9.3:** Consistency of MLE Estimators for Probit/Logit Determining Functions
- **Reasoning Box 9.4:** Confidence Intervals for Parameters of a Probit/Logit Determining Function
- **Reasoning Box 9.5:** Hypothesis Testing for Parameters of a Probit/Logit Determining Function

CHAPTER 10

Identification and Data Assessment

Data Challenge: Are Projected Profits over the Hill?

Introduction

Assessing Data Via Identification

Identification Problems and Remedies

Extrapolation and Interpolation

Variable Co-movement

Identification Damage Control: Signing The Bias

Rising to the Data Challenge

Summary / Key Terms and Concepts / Conceptual Questions / Quantitative Problems

- Communicating Data 10.1: Projecting Trends
- Communicating Data 10.2: Disentangling Promotion from Financing
- Communicating Data 10.3: A Distorted View of a Degree's Value
- Reasoning Box 10.1: Can Data Deliver the (Sufficiently Precise) Answer?
- Reasoning Box 10.2: The Effects of Variable Co-Movement on Identification
- Reasoning Box 10.3: Signing Omitted Variable Bias

xviii

APPLICATIONS

Data Analysis Critiques, Write-ups, and Projects

Introduction

Critical Analysis of Data-Driven Conclusions

Case 1: Tennis Analytics

Case 2: Switching Insurance

Case 3: Grocery Store Price Promotions

Written Explanations of Data Analysis and Active Predictions

Case 1: Insurance Claims and Deductibles

Case 2: Wearable Features and Sales

Case 3: Ad Duration and Clicks

Projects: Combining Analysis with Reason-Based Communication

Project 1: Tablet Price and Profits

Project 2: Auto Ad Budget and Revenues

Project 3: Machine Maintenance and Quality

Glossary

Index

The Roles of Data and Predictive Analytics in Business

1

LEARNING OBJECTIVES

After completing this chapter, you will be able to:

- LO1.1** Explain how predictive analytics can help in business strategy formulation.
- LO1.2** Distinguish structured from unstructured data.
- LO1.3** Differentiate units of observation.
- LO1.4** Outline a data-generating process.
- LO1.5** Describe the primary ways that data analysis is used to aid business performance.
- LO1.6** Discriminate between lead and lag information.
- LO1.7** Discriminate between active and passive prediction.
- LO1.8** Recognize questions pertaining to business strategy that may utilize (active) predictive analytics.

Chapter opener image credit: ©naqiewei/Getty Images

dataCHALLENGE *Navigating a Data Dump*

As a newly hired analyst at Papa John's, your boss comes to you with a data question. She notes that the company's database contains data collected from Facebook, the company website, and internal sources. Below is a list of variables from each source for which the company has data.

FACEBOOK	WEBSITE	INTERNAL SOURCES
Number of likes	Number of visits	Product sales
Age of likers	Number of pages per visit	Employee salaries
Education of likers	Bounce rate	Product prices
Income of likers	Location of visitors	Advertising expenditures
Click-through rate to website	Device used by visitors	Production costs

Your boss is interested to know if and how Facebook "Likes" lead to sales. The question your boss has posed is clear enough. However, she provides no guidance beyond the question.

How should you begin?

2

Introduction

Traditionally, data analysis was reserved for a small corner of the business community, performed by quantitatively oriented experts, if it was done at all. With the increasing digitization of business information, computer and software power, and information availability, data analysis has become an integral component of business. Data analysis can help businesses improve production

processes, customer marketing, and strategic positioning. The enormous list of applications makes it impractical to learn about all of them in just one book or one course.

Why should our focus be on *predictive analytics* for business strategy? First, the topic is relevant for virtually every business discipline. Every discipline must regularly formulate and evaluate strategic options, and all are affected by strategies undertaken by various branches of the firm. Second, predicting the consequences of a strategic move stems from perhaps the most fundamental application of data analysis: generating knowledge about cause and effect. The use of data to learn about cause and effect spans well beyond the business environment into areas such as medicine, physics, chemistry, and public policy, to name just a few. The critical thinking skills one must acquire to understand predictive analytics for business strategy have value well beyond this one area of study.

The first section of the chapter defines data and provides a concise characterization of predictive analytics for business strategy—one of many ways businesses use data. The next section details data features. For those already familiar with data, the data definition and many of the data features will be familiar; however, a clear understanding of predictive analytics for business strategy and a data-generating process are crucial for what follows in the book, and therefore these two sections merit a review. The third section gives a broad overview of data uses in business. The final section describes different temporal uses of data, i.e., measuring the past and present versus predicting the future, and makes a crucial distinction between two different types of prediction, passive and active.

The objective of this chapter is to paint a basic picture of the business analytics landscape and explain where predictive analytics for business strategy lives within that landscape. In doing so, we provide a general description of what this strand of business analytics does, and differentiate it from other types of business analytics and predictive analytics.

Defining Data and Data Uses in Business

We begin by defining data and other basic terminology relating to data. (Be sure to read [Communicating Data 1.1](#), which addresses the question of whether data is/are singular or plural.) We then characterize predictive analytics for business strategy by breaking it down into its core components.

3

DATA

Put simply, **data** are a collection of information. Often, we associate data with numerical entries in a table, stored in a computer spreadsheet. However, data are stored and organized in many other ways as well. The pencil marks that recorded your height on a wall during your childhood constitute a collection of data. The websites you visited each day over the past week that are logged in your computer's "History" folder are a collection of data.

data A collection of information.

Organized collections of data that firms use for analysis are called databases. A **database** consists of one or more tables of information, although some alternative database systems use other methods of storage, including domains. All the analytical methods we will discuss in this text use a single table of data; however, we will at times consider alternative methods that use multiple tables for comparison purposes. If the data to be analyzed are stored across several tables (or domains) within a database, it is a straightforward task to construct the appropriate single table/spreadsheet for analysis.

database Organized collection of data that firms use for analysis.

Fortunately, even when data are stored using other forms of media (pen and paper, marks on a wall), it is generally feasible to import them into a spreadsheet for analysis. Suppose you decided to keep track of how much you spent on lunch over the past five days. To do this, you may have simply

written down the dates and expenditures on a piece of paper as follows: 6/2/18, \$7.34; 6/3/18, \$8.42; 6/4/18, \$7.63; 6/5/18, \$5.40; 6/6/18, \$9.30. You could import these data to a spreadsheet as shown in [Table 1.1](#).

Putting the data into a spreadsheet often makes even the most basic analysis simpler. For example, we can see that the lunch expense on 6/5 was several dollars less than the others.

PREDICTIVE ANALYTICS WITHIN BUSINESS ANALYTICS

Once you've recorded some data, even data as simple as your lunch expenses for the week, you'll be able to analyze the data for any information they can provide. The use of data analysis to aid in business decision making is commonly referred to as **business analytics**. Business analytics is a broad field encompassing many (often overlapping) approaches to data analysis including econometrics, data mining, and predictive analytics.

business analytics The use of data analysis to aid in business decision making.

Predictive analytics is any use of data analysis designed to form predictions about future, or unknown, events or outcomes. Formulating accurate predictions for the future has obvious value in many contexts. Areas where predictive analytics are practiced are numerous and include fraud detection (predicting false insurance claims), risk management (predicting portfolio performance), and business strategy.

predictive analytics The use of data analysis designed to form predictions about future, or unknown, events or outcomes.

TABLE 1.1 Lunch Expense Data

DATE	EXPENSE
6/2/18	\$7.34

6/3/18	\$8.42
6/4/18	\$7.63
6/5/18	\$5.40
6/6/18	\$9.30

COMMUNICATING DATA 1.1

IS/ARE DATA SINGULAR OR PLURAL?

For people in business who use and talk data, the term's Latin roots can be cause for confusion regarding whether the word *data* is singular or plural. The answer is it can be either, depending on the context. Going back to the word's Latin origins, *data* is the plural of *datum* (meaning a piece of information). But although plural in Latin, *data* as an English word is treated both as a count noun (multiple items that can be counted) and as a mass noun (an entity with an amount that cannot naturally be counted). Hence, those who treat data as a count noun will say "The data are, or indicate" and those who treat data as a mass noun will say "The data is, or indicates."

Some data-oriented fields are particular about how to "properly" use the word. For example, in economics, there is a tendency to use the plural. Other fields tend toward the singular. Beyond this, it is a matter of preference. As an economist and former Latin scholar, I will use the plural version of data throughout this book; however, this is not to indicate it is more correct than the oft-used singular version.

BUSINESS STRATEGY

A **business strategy** is a plan of action designed by a business practitioner to achieve a business objective. Examples of business objectives include profit maximization, improved customer retention, and enhanced employee satisfaction. Examples of plans of action include pricing decisions,

advertising campaigns, and methods of employee compensation. A simple business strategy, for example, may be to lower a product's price in order to grow sales.¹

business strategy A plan of action designed by a business practitioner to achieve a business objective.

To remain competitive, managers and other decision makers must constantly be mindful of the business strategies they are employing and of alternative strategies that might be worth considering. While simply measuring corresponding outcomes may seem sufficient for evaluating strategies that are currently being followed, the crucial question is what one should expect a strategy to accomplish in the future. This expectation should be compared against expectations for competing strategies, so the manager can follow the optimal business strategy moving forward. The ability to make comparisons regarding the future depends on the ability to form reliable predictions, often using data.

PREDICTIVE ANALYTICS FOR BUSINESS STRATEGY

LO 1.1 Explain how predictive analytics can help in business strategy formulation.

Without data, business decision makers would have to rely on theoretical arguments alone to predict the effects of alternative strategies and ultimately decide which one(s) to enact. Some theoretical arguments are formal; others are based on “gut feelings” or “instinct.”

Although theoretical arguments can be compelling, they can become much more powerful and convincing when supported by data. For example, it is debated whether drinking diet soda helps people lose weight or actually causes them to gain weight. There are theoretical arguments on both sides. Those claiming it helps people lose weight argue that diet soda has fewer

calories, and fewer calories translates into lower weight. Those

5

claiming it causes weight gain argue that diet soda's inability to quench the body's craving for sugar creates even stronger cravings that will be satisfied by consuming unhealthy foods. Which claim is correct? With no supporting data, the decision becomes one based on "gut feeling." Now, suppose 1,000 individuals were split into two groups of 500: One group (Group A) was allowed to drink only regular soda, and the other group (Group B) was allowed to drink only diet soda. Then, after one year, suppose Group B showed a much greater increase in weight than Group A. Which prediction concerning diet soda would you believe after these data came in?

Predictive analytics is an ideal complement to the formulation of business strategy. It allows the decision maker to make evidence-based assessments (where the evidence consists of data) of expected outcomes from alternative strategies, and then choose the optimal one based on her or his business objective. Consider a manager who is trying to determine whether to change the price of a product, and if so, in what direction. The objective is to increase revenue in the short run. Gut feeling may suggest that a price increase will accomplish this task. However, before acting on this instinct, the manager collects data on prices and sales for the product, and then uses these data to estimate a predictive analytics model. The finding is that the price elasticity of the product is -2.1 . This means that a 1% increase in price leads to a 2.1% decrease in quantity demanded. Since price increases result in decreased revenue when price elasticity is greater than one (in absolute value), the manager decides that a price *decrease* is the proper strategy to follow based on this analysis.

Predictive analytics models do not always yield perfect predictions. These models rely on assumptions, some testable and others not. However, they provide a structured mechanism that allows us to use what actually has occurred (recorded as data) to inform us about what alternative strategies will accomplish. Understanding the level of reliability of the predictive analytical model being used is a crucial part of this process, and a recurrent theme in this book.

Data Features

LO 1.2 Distinguish structured from unstructured data.

Data possess several important features that affect their ability to be analyzed and the analysis method to be used. In this section, we discuss three key features that we should identify before attempting to analyze any dataset: whether the data are structured or not, the unit of observation, and the data-generating process.

STRUCTURED VS. UNSTRUCTURED DATA

The analytical methods we discuss in this book all apply to data of a particular type: structured data. Structured data are the type of data with which most casual observers are familiar; they are the data that come in a spreadsheet format. More formally, **structured data** have well-defined units of observation for which corresponding information is identifiable. By **unit of observation**, we mean the entity for which information has been collected. For example, the data in [Table 1.2](#) are structured data on yearly sales; the unit of observation is a year and the corresponding information is number of sales.

structured data Data with well-defined units of observation for which corresponding information is identifiable; they are the data that come in a spreadsheet format.

unit of observation The entity for which information has been collected.

Structured data need not be in spreadsheet format. Suppose you found a piece of paper in your desk, and at the top was written, “Yearly Sales.” Then, written haphazardly

YEAR	SALES
2012	4,382
2013	4,615
2014	4,184
2015	5,043
2016	5,218
2017	5,133
2018	5,391

across the paper you saw the following figures: 4,382(2012); 4,615(2013); 4,184(2014); 5,043(2015); 5,218(2016); 5,133(2017); 5,391(2018). Although not in spreadsheet format, these data are structured. As in the spreadsheet, the unit of observation is well defined (a year), and the corresponding information (sales) is identifiable.

Unstructured data are the complement to structured data; they are any data that cannot be classified as structured. A series of images is an example of unstructured data. The words in this paragraph are unstructured data. The data in [Table 1.3](#) are unstructured; this is because, although it seems clear these are yearly data, there is no way, with the information given, to identify the corresponding sales information for each year.

unstructured data Any data that cannot be classified as structured.

THE UNIT OF OBSERVATION

LO 1.3 Differentiate units of observation.

The crucial component of structured data is the unit of observation; we cannot build or analyze a structured dataset without it. In addition, the unit of observation tells us the way in which the information in the data varies. Does the information in the data vary across people, countries, time, people and time, etc.? As we will see in later chapters, understanding how the information varies in the data can be a critical factor in choosing and

assessing a method of analysis.

Determining the unit of observation essentially boils down to answering these four fundamental questions: What? Where? Who? When? The unit of observation becomes more refined as the data give answers to more of these questions, provided an answer (a) is not constant across all observations and (b) is not perfectly determined by the answer to another question.

For example, if the data consist of production figures for 10 factories in 2017, the unit of observation is a factory. In this case, the answer to “What?” is the factory, and there is variation

TABLE 1.3 Example of Unstructured Data

YEAR	SALES	2012
4,615	4,184	
2013	2016	5,133
5,043		2015
		4,382
5,391	5,218	
2017		
2018		2014

7

in this answer. There is no answer in the data to “Where?” and “Who?” and although there is an answer to “When?” there is no variation in this answer (it is always 2017).

Suppose that the data also provided the location of each factory. Despite this information, the unit of observation is still a factory, assuming no factories are able to change location. Here, knowing which factory you are observing perfectly determines the location as well; there is no refinement in the unit of observation.

When characterizing units of observation, the nature of the time component of the data often gets special attention, because it can ultimately affect the proper method of analysis. The four main groupings of units of observation are cross-sectional data, pooled cross-sectional data, time-series data, and panel data.

Cross-sectional Data Cross-sectional data exhibit no variation in time. When determining the unit of observation, the answer to “When?” is constant. Consequently, cross-sectional data provide a snapshot of information at one fixed point in time. The point in time need not be short in duration; it can be an hour, a day, a month, or even a year. Examples of cross-sectional data include the following:

cross-sectional data Data that provide a snapshot of information at one fixed point in time.

- Sales made by each employee on July 12, 2017.
- Visits to the top 100 websites in January 2018.

1.1

Demonstration Problem

Your IT group just informed you that it has a new dataset on employee salaries. Before conducting analysis on this new dataset, you want to establish the unit of observation. Consequently, you start working through the four “W” questions:

- *Who*: Given these are data on employees, the natural first “W” is “Who?” In this case the answer is employees, and as long as there is more than one employee, this is a relevant dimension in determining the unit of observation.
- *What*: The next “W” might be “What?” A possible answer may be a company's division, as an employee may move around within the firm and even within location. For this example, let's suppose there is no variation of this sort, so answering “What?” provides no further refinement.
- *When*: Next, we might ask “When?” As these are salaries, the answer to this question would be the year of the observation. If employees are observed over multiple years, this is also a relevant dimension in determining the unit of observation.

- *Where*: Lastly, let's ask "Where?" A possible answer may be the location of the firm branch where the employee worked. If at least some employees worked at multiple locations in the same year, then this also would provide a relevant dimension in determining the unit of observation.

Now, suppose for this particular dataset, your series of "W" questions yielded: (Who), Employees; (What), Not applicable; (When), Year; and (Where), Determined by Year. Let's say the last answer you get indicates that employees did not switch locations within a year, and so knowing the location provides no further refinement beyond knowing the year. Consequently, the unit of observation is employee-year; that is, a new observation entails a change in the employee, the year, or both.

TABLE 1.4 Example of Cross-sectional Data

EMPLOYEE	SALES	DATE
Herbert McDunnough	12	7/12/17
Carl Showalter	10	7/12/17
Marge Gunderson	19	7/12/17
Loretta Bell	7	7/12/17
Walter Sobchak	15	7/12/17
Mattie Ross	12	7/12/17

- Gross domestic products for every country in 2017.
- Height and weight of 5,000 randomly selected individuals in the United States in May 2017.

We illustrate the first of these examples in [Table 1.4](#).

Pooled Cross-sectional Data Cross-sectional data may be collected more than once. One might have collected the height and weight of 5,000 randomly selected individuals in the United States in May 2017 and then the same

information for another random sample of 5,000 in May 2018. There is no systematic relationship between the units of observation in the first sample and those in the second sample. Of course, it is possible that the same person could be included in both samples, but this would be purely by chance, not by design.

For some analyses, it is beneficial or even crucial to combine two cross-sections into one dataset. In our height and weight example, the new dataset is no longer a cross-sectional dataset, because there is now variation in “When?” (2017 *and* 2018). When we combine two or more unrelated cross-sectional datasets into one dataset, the result is **pooled cross-sectional data**. An illustration of the height/weight example is in [Table 1.5](#).

pooled cross-sectional data The result of two or more unrelated cross-sectional datasets being combined into one dataset.

Time-series Data **Time-series data** exhibit only variation in time. For these data, the answers to “What?” “Where?” and “Who?” do not change across observations, but the answer to “When?” does. Many macroeconomic datasets are time-series data. Examples include annual gross domestic product for the United States from 1950 to 2018; monthly Indian interest rates from 2000 to 2017; and monthly unemployment rates in Mexico from 1990 to 2015. Businesses also collect some time-series data. Examples include annual firm profits, 1995–2018; annual CEO compensation, 1991–2016; and monthly employee turnover, 1999–2017. An illustration for firm profits is in [Table 1.6](#).

time-series data Data that exhibit only variation in time.

TABLE 1.5 Example of Pooled Cross-sectional Data

NAME	HEIGHT (INCHES)	WEIGHT (POUNDS)	MONTH
Angela Hoenikker	56	138	May 2017
Dwayne Hoover	59	172	May 2017
.	.	.	.

Paul Lazzaro	63	195	May 2018
Valencia Merble	58	152	May 2018

TABLE 1.6 Example of Time Series Data

YEAR	PROFITS (MILLIONS)
1995	\$4.3
1996	\$5.2
.	.
.	.
2017	-\$2.2
2018	\$1.4

Panel Data In business, many datasets involve observing the *same* cross-sectional units over multiple points in time. Such data are called **panel data**. Panel data look very similar to pooled cross-sectional data; the only difference is that, for panel data, the cross-sectional units in the dataset are the same across time, whereas for pooled cross-sections, they generally are not. Examples of panel data include sales made by a select group of employees on both July 12, 2017 and August 12, 2017; monthly visits to Yahoo, Google, and Bing from January 2015 to December 2017; and annual gross domestic product (GDP) for every country from 2014 to 2017. The search engine example is illustrated in [Table 1.7](#).

panel data The same cross-sectional units over multiple points in time.

We conclude our discussion on data types with [Communicating Data 1.2](#), in which we provide some practical examples of how to explain data features.

DATA-GENERATING PROCESS

LO 1.4 Outline a data-generating process.

A fundamental question to ask when confronting a new dataset is: “How did these data come about?” Put another way: “What is the data-generating process?” The **data-generating process** (DGP) is the underlying mechanism that produces the pieces of information contained in a dataset. Performing data analysis without an understanding of the DGP is analogous to assessing college quality using a third-party ranking without understanding how the ranking was done. For example, if one ranking is completely based on average local temperature and another is based solely on average rate of employment after graduation, each has a very different interpretation with regard to quality.

data-generating process (DGP) The underlying mechanism that produces the pieces of information contained in a dataset.

TABLE 1.7 Example of Panel Data

MONTH	SEARCH ENGINE	VISITS (MILLIONS)
January 2015	Yahoo	147
January 2015	Google	183
January 2015	Bing	112
.	.	
.	.	
December 2017	Yahoo	171
December 2017	Google	205
December 2017	Bing	148

COMMUNICATING DATA 1.2

ELABORATING ON DATA TYPES

When you describe a dataset, it is expedient to simply state the data type (e.g.,

cross-sectional). However, many in the business world, even those who might be making data-based decisions, don't have a ready understanding of these terms. Consequently, there is value in developing a clear, accurate way of explaining these data features. Consider the following examples, in which both a general and a specific (with hypothetical details) response are possible:

- A. You have data on branch-level employee turnover in the form of a panel.
What does this mean?

Possible Answer:

(General) I observe employee turnover for each branch of the company on several different occasions.

(Specific) I observe each branch's turnover each month from January 2015 to December 2016.

- B. You have data on customer satisfaction in the form of a pooled cross-section. What does this mean?

Possible Answer:

(General) I observe customer satisfaction for several different groups of customers, each group asked at a different point in time.

(Specific) I observe three different groups of customers, each consisting of about 1,000 people; the first in 2014, the second in 2015, and the third in 2016.

- C. You have data on one restaurant's pricing in the form of a time series.
What does this mean?

Possible Answer:

(General) I observe prices for a single restaurant at many different points in time.

(Specific) I observe prices for a single restaurant each week from January 4, 2016, until June 4, 2016.

- D. You have cross-sectional data on county population. What does this mean?

Possible Answer:

(General) I observe county-level population for many counties at a single point in time.

(Specific) I observe county-level population for all continental U.S. states on July 31, 2016.

Establishing the data-generating process can be both informal and formal. Informally, it involves considering all of the factors that together determine individual observations. For example, a student's grade point average depends on her innate ability, the time she spends studying, the quality of teaching she receives, and her health throughout the term, among other things. This is an informal description of the data-generating process for grade point averages in a dataset.

The process of informally establishing the data-generating process may seem quite basic, and in fact, it is. However, this process can be highly useful, especially at the beginning of data analysis. In many industries, it is common to be confronted with large volumes of data on a wide range of variables, often with several sources. The analytical possibilities can be endless, and so can be the relationships (correlations) that one can find among these variables. Developing an informal concept of the data-generating process can help to sort through which of these relationships is worth exploring, or even makes sense.

1.2

Demonstration Problem

Let Y be the starting salary for a given person's first job. This is a random variable that can take on many different values (e.g., \$20,000, \$45,000, \$82,000). Suppose you want to informally establish the data-generating process for Y . What are factors that likely contribute to the realized values for Y ? (Try this on your own before reading further.)

The possible contributors to Y 's realized value clearly include age, level of education, industry, location, and the unemployment rate. Can you come up with others?

To illustrate this point, consider a firm that has weekly data on the number and originating location of visits to its website, as well as productivity levels (i.e., output per unit of labor) for one of its factories in Bangalore, India. An analyst could determine the correlation in the number of visits to the website from the Northeast United States and productivity levels in Bangalore. However, what meaning can be drawn from this correlation? Does productivity depend on these website visits, or vice versa? Or, is there another variable upon which both depend? Without some establishment of the data-generating process, it is difficult to find use with this measure, or to justify taking the time to collect the data. Clearly, before you decide to collect and analyze data, you need to spend some time thinking what variables will be meaningful.

Formally establishing a data-generating process involves building a representative statistical model. Such models typically treat the components of a dataset as realizations of random variables. For example, if the attendance at an amusement park depends entirely on the temperature that day, a formal representation of the data-generating process may look like:

$$\text{Attendance}_t = f(\text{Temperature}_t)$$

In words, this means that the attendance at the amusement park on day t is a function of the temperature on day t .

Of course, amusement park attendance on a given day depends on many other things beyond that day's temperature. Perhaps the simplest way to account for such additional factors is to extend the above model as follows:

$$\text{Attendance}_t = f(\text{Temperature}_t) + U_t$$

Here, we can think of U_t as “all other factors, besides temperature, affecting amusement park attendance.”

Building a formal model of the data-generating process need not be more complex than this simple example for amusement parks. It need only provide sufficient structure upon which meaningful analysis can be conducted. While it is not crucial to establish a formal model at the onset of the analytical process, it becomes critically important later on. In fact, every analytical technique discussed in this book utilizes a formal (often simple) model of the data-generating process.

12

Basic Uses of Data Analysis for Business

LO 1.5 Describe the primary ways that data analysis is used to aid business performance.

Data analysis has many uses beyond predictive analytics for business strategy. Before diving into the specific focus of this book, it is useful to have some perspective on the range of applications of data analysis in the business world. We can categorize business uses of data analysis into the following categories: queries, pattern discovery, and causal inference. These categories do not encompass all possible uses of data for business purposes; however, they cover a large proportion and provide an intuitive way to begin sorting through this enormous field.

QUERIES

Possibly the most basic and ubiquitous use of data analysis in business (and virtually every other field) is for the purpose of information retrieval and summary. Even relatively small datasets contain very large amounts of information, consisting of the information in the observations themselves, as well as information resulting from combinations of the observations. Sorting

through all of this information essentially boils down to asking questions of the data (e.g., “What are average profits?” “Who had the most sales last quarter?”). Any request for information from a database is called a **query**.

query Any request for information from a database.

Many software packages are designed to streamline the asking and answering of queries via a database. Even for analytics novices, the use of query software is likely quite familiar. Anyone who has used a search engine on the Internet already has experience with query software. Suppose you entered “Analytics textbooks” in the Google search bar. Google treats the universe of web pages as a giant database, and then treats this search as though you entered this query: “What are the most relevant web pages for analytics textbooks?” The concept of “relevant” is not particularly well defined here; each search engine formalizes relevance in its own way (e.g., a combination of the number of times the search words appear on the page, number of links to the page, etc.). Then, the search simply answers the query with a ranking of web pages, starting with the most relevant. Every search you conduct on Google is a query for the database of web pages.

As a simple example of queries for a firm's database, consider the data in **Table 1.8**. These are sales data for three employees, who make sales in two different locations. When analyzing these data, you may want answers to the questions “Which city had the most sales?” or “Which employee had the most sales in Chicago?” or “How many sales did Robert Jordan make in New York?” These are all examples of queries. With such a small dataset, we can answer these queries by simply looking at the data (Chicago, Catherine

TABLE 1.8 Dataset of Sales for Three Employees

EMPLOYEE	LOCATION	SALES
Robert Jordan	New York	87
Robert Jordan	Chicago	63
Catherine Barkley	New York	78
Catherine Barkley	Chicago	91
Bill Gorton	New York	57

Barkley, and 87, respectively). Of course, with larger datasets, this “eyeball” approach is not possible. Fortunately software is available (just like Google’s search engine) that can comb through data to find the answers to a wide array of queries.

A class of queries worthy of special mention is descriptive statistics, which are a standard component of virtually any series of queries in practice. **Descriptive statistics** are broadly defined as quantitative measures meant to summarize and interpret properties of a dataset. They typically consist of measures designed to assess the “center” and “spread” of variables within a dataset. For analyzing the dataset contained in [Table 1.8](#), the mean and variance of sales are both examples of descriptive statistics—the former serving to locate the “center,” and the latter serving as a measure of “spread.” Here, the mean of sales is 74.5 and the variance is 178.3.

descriptive statistics Quantitative measures meant to summarize and interpret properties of a dataset.

While descriptive statistics are standard queries for data analysis, we can pose many other queries. Even for small datasets, the number of queries we can construct is exceedingly large. This raises the question of how to choose appropriate queries beyond simple descriptive statistics. In business and elsewhere, a primary motivator in choosing queries is performance evaluation. Often just a few queries can provide valuable insights as to how well an individual, team, firm, or even country is performing.

One of the most well-known and long-standing applications of using queries for performance evaluation outside of business is in the sport of baseball. Throughout the history of the game, players have been defined by their batting average, number of home runs, number of stolen bases, earned run average, and so on. Owners, managers, and fans have used many such statistics to evaluate how good players are. As data have become more plentiful in baseball, the number and types of queries that can be answered

have increased dramatically. A manager may want to know a player's batting average against left-handed pitching, with two outs, after the fifth inning, with the bases loaded, in close games (e.g., when the teams are within three runs of each other), and in other situations. This lets a manager evaluate whether the player is a good choice to bat in a given situation, or if there is another player on the roster who might be better. We provide more discussion on baseball queries in [Communicating Data 1.3](#).

Queries are pervasive in business. Simple descriptive statistics such as average profit per sale, average sales per store, the variance in sales across stores, and average employee health care costs can all serve as important and insightful measures for performance evaluation. Which age group spends the most on our product? What were production levels in Europe last month? How long does the average employee stay with our firm? What is the trend in profits over the past six months? Which salesperson posted the most sales last year? Queries such as these can be highly valuable.

In many cases, a manager will determine a class of queries (e.g., total sales for each employee), and instead of answering them one at a time as separate questions, will prefer to look at a summary of the data to quickly answer and explore many queries at once. Then, a pivot table becomes a valuable tool. A **pivot table** is a data summarization tool that allows for different views of a given dataset. These tools are often used in spreadsheet software, such as Excel.

pivot table A data summarization tool that allows for different views of a given dataset.

To build a pivot table, you need only determine the desired measure and filtering dimensions. The measure is represented as a number, while the dimensions can be numbers or text. For example, consider a cross-sectional dataset where each observation contains information on the following: sales, location, price, and salesperson. In building a pivot table, you may choose your measure to be total sales, and a dimension to be location.

COMMUNICATING DATA 1.3

SITUATIONAL BATTING AVERAGES

Some queries in baseball can be particularly informative for team managers and owners. For example, teams consistently query batting data to determine how their players perform differently depending on whether runners are on base. A high batting average with runners on base more readily translates into runs, and ultimately wins.

Consider two prominent Major League batters, Mike Trout and Bryce Harper. Looking at the data in [Table 1.9](#), you can see the performance change for the players in each situation during the 2016 season. Harper's batting average improves significantly when there are runners on base as opposed to when there is no one on base. Mike Trout's remains almost the same in both situations.

TABLE 1.9 Situational Batting Average Data

PLAYER	RUNNERS ON	NONE ON
Mike Trout	0.311	0.318
Bryce Harper	0.266	0.223

TABLE 1.10 Pivot Table with Sales by Location

LOCATION	TOTAL SALES
East	107
West	215
North	135
South	281
Grand total:	738

This choice would result in a pivot table that looks like [Table 1.10](#). Another possibility is to choose your measure as average price, and dimensions of location and salesperson. A pivot table with these choices appears in [Table](#)

1.11.

TABLE 1.11 Pivot Table with Price by Location and Salesperson

AVERAGE PRICE	SALESPERSON			
LOCATION	JILLIAN	THOMAS	ANN	GRAND TOTAL
East	\$21.85	\$22.16	\$24.31	\$22.81
West	24.86	27.19	28.14	\$26.68
North	20.18	24.91	22.35	\$22.27
South	23.84	21.18	24.71	\$23.67
Grand total:	\$22.73	\$24.41	\$24.27	\$23.80

15

PATTERN DISCOVERY

In addition to queries, businesses also use data analysis for the purpose of pattern discovery. A **pattern** in a dataset is any distinctive relationship between observations within the dataset. **Pattern discovery** is the process of identifying distinctive relationships between observations in a dataset. Pattern discovery can be synonymous with **data mining**; however, data mining typically involves pattern discovery in large datasets. Data mining is a subset of pattern discovery.

pattern Any distinctive relationship between observations within the dataset.

pattern discovery The process of identifying distinctive relationships between observations in a dataset.

data mining Pattern discovery, typically in large datasets.

The definition of a pattern is quite general, and for good reason. A wide array of relationships may be worth discovering, depending on the purpose of the analysis. For example, supposing each observation contains information about variable X (e.g., price) and Y (e.g., sales), then the correlation between

X and Y may be a pattern that is discovered in the data. *Linear regression*, in the context of pattern discovery, is another method of discovery that looks for partial correlations between variables. *Classification* looks for analogs to partial correlations between variables when at least one of the variables is categorical. We'll discuss linear regression in great detail throughout the book, and we will further discuss classification later in this chapter (in the section Passive Prediction).

Other examples of pattern discovery include association analysis, cluster analysis, and outlier detection. **Association analysis** attempts to discover dependencies (generally in the form of conditional probabilities) between two or more variables in the data. **Cluster analysis** groups observations according to some measure of similarity, so observations in the same group are more similar than observations in different groups. **Outlier detection** finds small subsets of observations (if they exist) that contain information far different from the vast majority of the observations in the dataset.

association analysis Attempts to discover dependencies, generally in the form of conditional probabilities, between two or more variables in the data.

cluster analysis Groups observations according to some measure of similarity.

outlier detection Small subsets of observations, if they exist, that contain information far different from the vast majority of the observations in the dataset.

To see some of these patterns in practice, consider the data presented in [Table 1.12](#). Assume this table consists of a sample of individuals that shopped for product X on that product's website in a given month. Here, the dataset is small enough that we can discover patterns like those described above through simple formulas or even by just looking at the data. For larger datasets, we generally need data mining software to accomplish these tasks.

In [Table 1.12](#), we can discover the following patterns. First, individual #7

is a clear outlier both in the fact that his income is far above anyone else's, and the age/income combination is highly unusual (young and wealthy, unlike anyone else in the sample). The correlation between age and income for the entire dataset is -0.0909 . However, after removing the outlier (individual #7), the correlation is 0.9475 . This shows how important

TABLE 1.12 Website Shopping Data

OBSERVATION NO.	AGE	INCOME	PURCHASE
1	53	\$110,000	Yes
2	24	38,000	No
3	42	80,000	Yes
4	60	95,000	Yes
5	28	41,000	Yes
6	30	44,000	No
7	27	430,000	Yes
8	62	140,000	No

16

outlier detection can be, and that the general relationship (for all but one observation) between income and age in these data is a strong, positive one.

In applying association analysis to the data in [Table 1.12](#), we can establish the following association rule: A purchase is associated with income of at least \$80,000. This stems from the fact that, conditional on having income of at least \$80,000, the probability of a purchase is 80% (notably higher than the unconditional probability of a purchase).

Lastly, these data suggest a clustering according to income and age. We can cluster individuals 2, 5 and 6 in one group (younger with low income) and individuals 1, 4 and 8 in another group (older with higher income).

In some instances, pattern discovery may seem no different from a query. Determining the correlation between two variables per se is really just a query of the data (“What is the correlation between X and Y ?”). To conclude that we’ve actually discovered a pattern in the dataset, we need some sense of what makes the relationship we’ve found “distinctive.” We can impose criteria, or a rule (or rules), to make this determination. A simple and

common way to do this is to establish a threshold. For example, we may require that correlation between two variables qualifies as a pattern only if it is larger than 0.4. Or, an association rule (B is associated with A) can be called a pattern only if the conditional probability is larger than 0.7. For the candidate patterns we described from [Table 1.12](#), the correlation (excluding the outlier) and the association rule we find satisfy these criteria, respectively. (Recall that the probability of a purchase given income of at least \$80,000 is 0.8.)

Pattern discovery for outlier detection and cluster analysis often involves thresholds for (Euclidean) distance. In determining whether the clusters (2,5,6) and (1,4,8) and outlier (7) we found qualify as a pattern, we can incorporate a distance threshold in Income/Age space. For the clusters, a simplified approach might involve setting a radius for a circle and assessing whether, say, three or more observations can fit within a circle of that size in the graph; when they can, that group is a cluster. For an outlier, we may set a radius for a circle and assess whether there are any observations that are the only ones contained in a circle of that size, when they are at the center of that circle. In [Figure 1.1](#), we plot all eight

FIGURE 1.1 Example of Outlier Detection and Cluster Analysis

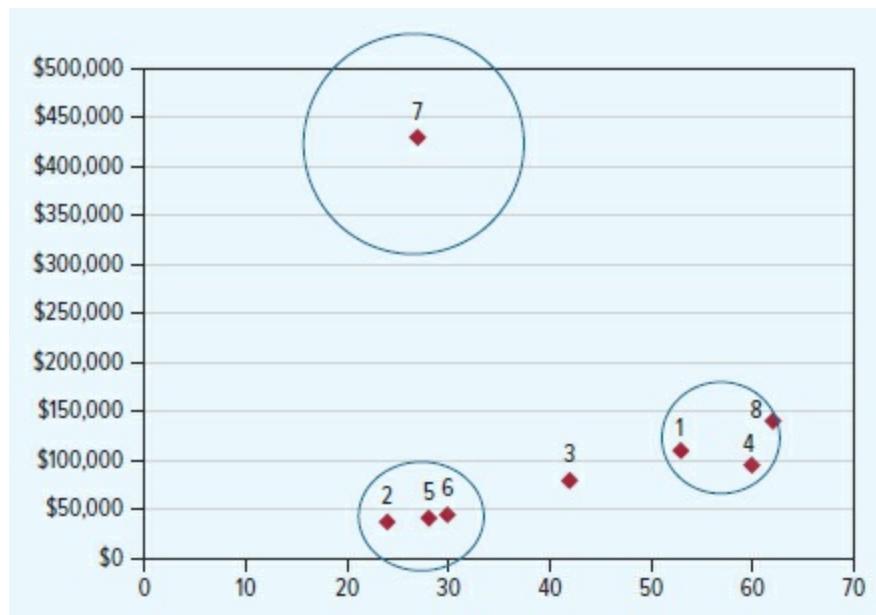
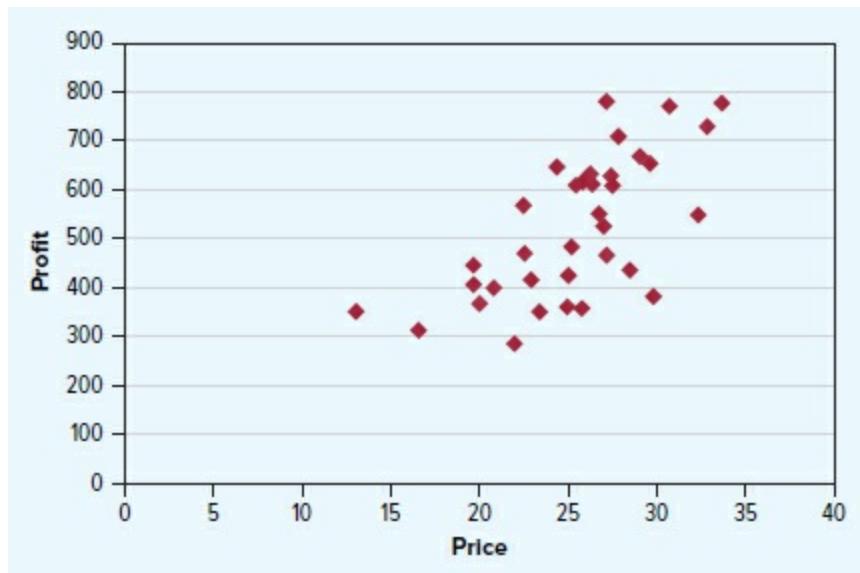


FIGURE 1.2 Scatterplot of Data on Profit and Price

observations. This plotting allows us to see the short distance between group members; these two groups are the only groups of three that can fit in the prescribed circle. It also allows us to see the long distance between each individual and individual #7 as illustrated by the fact that the circle around individual #7 would contain other individuals if it were drawn around any other point.

Pattern discovery has many applications in business. When focusing on the customer, pattern discovery helps firms construct customer segments, initiate segment-based promotional campaigns or even pricing, and form predictions about consumer behavior (will a customer default on a loan?). Business applications of pattern discovery go well beyond enhancing customer relations. Pattern discovery can assist a firm's human resources department to identify common characteristics among its most and least successful employees. Such discovery can help the firm build a more efficient system of employee recruitment, training, and retention.

Patterns that firms discover in data can even be highly suggestive of a causal relationship. Suppose an analyst has collected price and profit data

across many markets and plotted these data as in [Figure 1.2](#). What pattern seems most evident in this scatter plot? Here we discover that profits tend to be higher when price is higher. This observation can be useful toward understanding differences across markets, and it may be tempting to take it a bit further. The business may conclude from this pattern that raising price (e.g., from \$20 to \$30) would lead to an increase in profit. Of course, if the firm knew this to be true, it would be very useful information with regard to pricing strategy. However, simply discovering this pattern does not necessarily imply a causal relationship between these variables. Establishing causality requires the data analyst to approach (and sometimes construct) the data in quite particular ways.

CAUSAL INFERENCE

In addition to queries and pattern discovery, businesses use data analysis for causal inference. **Causal inference** is the process of establishing (and often measuring) a causal relationship between a variable(s) representing a cause and a variable(s) representing an effect,

causal inference The process of establishing (and often measuring) a causal relationship between a variable(s) representing a cause and a variable(s) representing an effect, where a change in the cause variable results in a change in the effect variable.

18

where a change in the cause variable results in a change in the effect variable. Consider two main types of causal relationship: direct causal relationships and indirect causal relationships. A **direct causal relationship** is one in which a change in the causal variable (X) directly causes a change in the effect variable (Y). For example, a person's caloric intake has a direct causal relationship with his weight. In contrast, an **indirect causal relationship** is one in which a change in X causes a change in Y , but only through its impact on a third variable. If daily exercise causes you to sleep better, which then causes

you to drink less coffee, we can say that daily exercise indirectly caused a reduction in coffee consumption. We provide a more detailed example of an indirect causal relationship in [Communicating Data 1.4](#). Throughout the book, we will focus our causal discussions on direct causality, except where otherwise noted.

direct causal relationship A change in the causal variable, X , directly causes a change in the effect variable, Y .

indirect causal relationship A change in X causes a change in Y , but only through its impact on a third variable.

Causal inference plays an important role in a wide range of fields in business and elsewhere. Examples of possible causal relationships that may be particularly important include: the effect of a drug on cancer remission, the effect of a tariff on imports, and whether combining two chemicals causes a chain reaction.

For business applications, causal inference typically takes place in one of two ways. The first is through experimentation. A firm may randomly fluctuate variables under its control in order to establish their effects on outcomes it values. By doing this, the firm is directly affecting the data-generating process since it is driving the mechanism that generates at least one of the variables. For example, suppose a popular website is wondering whether visitors are more likely to click on a banner ad if it is placed in the upper left corner vs. the upper right corner of the front page. Over several days, that website may randomly vary where the ad is placed for each new visitor, and then record the click-through rate for each placement. If the click-through rate for “upper left” is 0.05 and for “upper right” is 0.07, the website designers may conclude that the effect of moving an ad from “upper left” to “upper right” is an increase in the click-through rate of 0.02.

The second way causal inference typically takes place in business is through econometric models. Broadly speaking, *econometric models* are statistical models for economic and financial data. However, a primary objective of many of these models is the establishment of causality between

variables. Rather than directly influencing the data-generating

COMMUNICATING DATA 1.4

INDIRECT CAUSAL RELATIONSHIPS IN PURSE KNOCKOFFS

Indirect causal relationships can be found everywhere, including street vending in New York City. Consider the relationship between rainfall and sales of knockoff (imitation of designer) purses. Such purses are regularly offered for sale outside city parks and on the sidewalks by street vendors. It is reasonable to believe that sales of these purses will depend on the weather, but is this dependence direct or indirect?

While we could make a case for both types of dependence, there is almost certainly a large indirect causal relationship. Specifically, bad weather (heavy rain) will reduce the number of pedestrians, which will then reduce the number of purse sales for the street vendors. Rain does not directly affect someone's desire to purchase a purse, but it affects her desire to linger outdoors, which inhibits her from purchasing from one of the street vendors. Hence, weather affects sales indirectly through its effect on street traffic.

19

process as in experimentation, analysts can use econometric models, along with important assumptions about their properties, to model the data-generating process they observe. If the model and corresponding assumptions are correct, then estimating that model using the data can lead to causal inference. How this works (e.g., the process, reasoning) is a main focus of this book, and beyond what can be fully detailed in this chapter. For illustration purposes, however, consider the following example.

Suppose you have data on sales and prices for 2-liter bottles of Pepsi-cola for a large number of vendors. In this case, you are not directly influencing the data-generating process since you are only observing prices and sales rather than, for example, directly controlling the prices charged. A standard

measure of interest would be the elasticity of demand for this product. Given these data, you could easily calculate this measure by determining the average percentage change in units sold associated with a percentage change in price. However, an elasticity measure is causal in nature, in that it indicates the causal effect (in percentage terms) of a percentage change in price on units sold. To establish a causal relationship using these data requires you to model the data-generating process for prices and sales and make appropriate assumptions about the model.

Establishing causal relationships is highly important in many areas of business. Two particularly valuable areas of application are campaign evaluation and prediction. For *campaign evaluation*, a firm uses causal inference to determine the level of success or failure for a campaign it has undertaken. For example, if a firm engages in a promotional campaign involving promotional prices and advertising, it can use causal inference during and after the campaign to assess how well it performed with respect to profits, sales, etc. For *prediction*, a firm can use causal inference to predict the effect(s) of alternative strategies it may be considering. This application is the primary focus of the remainder of the book.

Data Analysis for the Past, Present, and Future

An important means of characterizing data analysis in business is according to the timing of the application. Is the analysis designed to assess what *happened*? what *is happening*? or what *is going to happen*? Establishing which of these questions is to be answered is key in determining the method(s) of analysis, presentation, and interpretation.

LAG AND LEAD INFORMATION

LO 1.6 Discriminate between lead and lag information.

From a timing perspective, data analysis can provide two different types of

information: lag information and lead information. **Lag information** is information about past outcomes. Lag information typically (but not always) contains information on variables classified as **key performance indicators (KPIs)** or variables that are used to help measure firm performance. Lag information is designed to answer the question, “What happened?” Often, firms want to know this information with very little delay, ideally in “real time.” For example, a firm may wish to know the rate at which people are visiting its website at a given point in time (e.g., measured as the number of visits in the past five minutes). Such information is lag information; however, if it is delivered with minimal delay (in seconds or less), it essentially can answer the question, “What is happening?”

lag information Information about past outcomes.

key performance indicators (KPIs) Variables that are used to help measure firm performance.

20

In contrast, **lead information** is information that provides insights about the future. It is designed to answer the question, “What is going to happen?” Lead information is generally used to help firms in the planning process, as they form expectations about the future and consider various strategic moves.

lead information Information that provides insights about the future.

Data analysis that is used for queries, pattern discovery, or causal inference can generate lag information. For example, a query asking which employee had the most sales over the past six months generates lag information that can be used for employee evaluation. With regard to pattern discovery, data analysis may lead to a discovery that web traffic for a firm's website is extremely low on Monday afternoons. This discovery of a segment of time as an outlier is another form of lag information and can be used as part of performance measurement for the website. For causal inference, a firm may use data analysis to establish the impact of a recent advertising

campaign on sales, i.e., how much did sales change as a result of the campaign? This lag information can be used to reward or penalize employees responsible for the campaign design and/or execution.

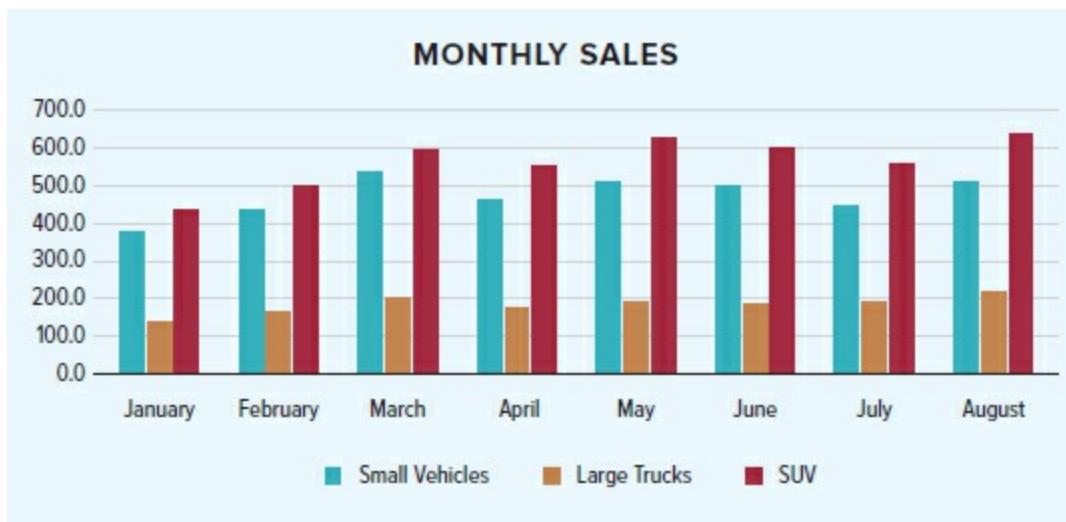
Businesses generate a great deal of lag information using standardized formats. Examples include reports, dashboards, and scorecards. These formats generally are used for presenting lag information for queries, often with built-in methods for performance evaluation. A **report** is the most broadly defined of these formats; it is any structured presentation of the information in a dataset. For example, the output presented in Excel when you construct a specific pivot table is a report. Reports are intended to make it easier to find and process desired information in a dataset. They often come in the form of a table, chart, or graph. [Figure 1.3](#) is an example of a report on monthly vehicle sales in the form of a chart.

report Any structured presentation of the information in a dataset.

A **dashboard** is a graphical presentation of the current standing and historical trends for variables of interest, typically KPIs. Dashboards derive their name from the fact that they are designed to provide some notion of the “speed” of the company, like a dashboard for a car. However, they typically do a bit more than this, since they often provide trend information, giving the analyst a sense of the firm's trajectory. Dashboards are used as a real-time monitoring device. By looking at a dashboard, the analyst should be able to establish how the firm is currently doing, and has been doing, with respect to one or more KPIs. [Figure 1.4](#) is an example of a dashboard, again for auto sales.

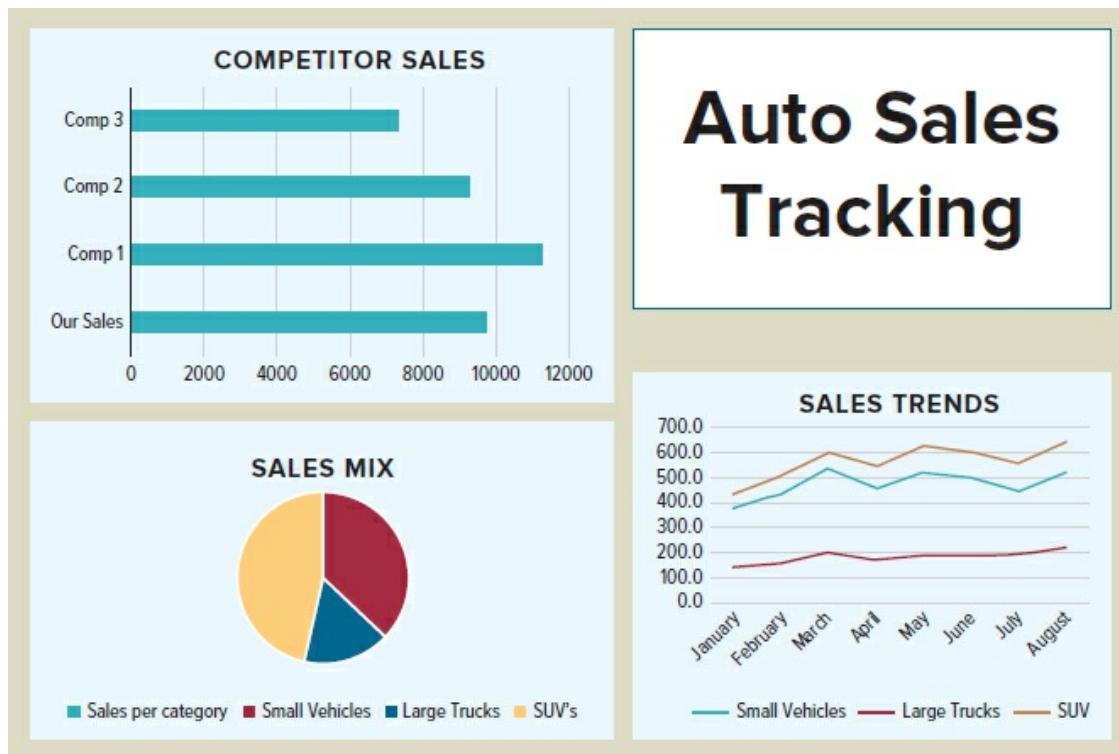
dashboard A graphical presentation of the current standing and historical trends for variables of interest, typically KPIs.

FIGURE 1.3 Report on Monthly Vehicle Sales



21

FIGURE 1.4 Dashboard for Auto Sales



A **scorecard** is any structured assessment of variables of interest, typically KPIs, against a given benchmark. Scorecards derive their name

from the fact that they help to “score” how the firm is performing along various dimensions. In contrast to dashboards which simply monitor KPIs, scorecards assess and communicate whether a given KPI level is good or bad, above average or below average, etc. It is up to the designer of the scorecard to determine the benchmarks, e.g., what levels of a given KPI should be considered “good” and what levels should be considered “bad.” [Figure 1.5](#) is a scorecard for a range of KPIs, again for an automobile firm.

scorecard Any structured assessment of variables of interest, typically KPIs, against a given benchmark.

Data analysis used for pattern discovery and causal inference, but typically not queries, can generate lead information. Recall that a query is any request for information from a database. These requests are not designed to predict the future but rather to answer questions about what is and what was. In contrast, discovery of a pattern can sometimes provide insights about the future. Suppose you discovered via data mining that visitors to your website who stayed longer than five minutes made a purchase 92% of the time. Then, if you observe a visitor to your website who has been there for seven minutes, what might you predict? It would be natural to predict that this person is highly likely (92% likely) to make a purchase.

Data analysis used for causal inference often can provide insights about the future as well. For example, your data analysis may identify a causal relationship between the price you charge and your sales, such that a 1% increase in price causes a 2% decrease in sales.

FIGURE 1.5 Scorecard for Automobile Firm

Goals	Measure	Follow Up	Target	Result	Performance	Initiative	Responsible
Financial	Associated Sales Costs	Monthly	\$250 per sale	\$500	Too High	Spend less time on paperwork with each customer	Sales Manager
Customer	Customer Experience Reviews	Weekly	3 out of 5 points	4	Good Level	Spend more time understanding customer needs before sales process	Sales Manager
Process	Time between delivery and moving vehicles to lot inventory	Monthly	3 days	2 days	Good Level	Redesigning detailing and unpacking process	Inventory Control Manager
Learning	Employee Vehicle Knowledge	Monthly	85% on manufacturer test	80%	Nearing Acceptable Level	New training program on vehicles and accessory packages	Sales Manager

Knowing this can provide insights about the effect of a future price change. You can use this to predict the effect on sales of a proposed 3% price hike: You would predict this price increase will lead to a 6% decline in sales.

Unlike lag information, lead information generally is not presented in some standardized format. Formulating predictions about the future requires more than just the information/data at hand; it typically requires accompanying assumptions and a line of reasoning, often along with a formal model of the data-generating process. Because of its greater complexity, we do not provide an example of lead information presentation here. Rather, in the Applications section at the end of the book, you will have ample opportunities to present lead information in the form of causal analysis using the information presented in the upcoming chapters.

PREDICTIVE ANALYTICS

LO 1.7 Discriminate between active and passive prediction.

Recall that predictive analytics is any use of data analysis designed to form predictions about future, or unknown, events or outcomes. Also recall that lead information is information that provides insights about the future. Consequently, predictive analytics is data

1.3

Demonstration Problem

Characterize the following types of information as lead or lag information:

1. Information on which customers are most likely to drop their cable service.
2. The probability that a given employee will leave the company in the next year.
3. Sales growth in each region during the most recent six months.
4. Expected number of Facebook Likes for a new product rollout.
5. Managers who engaged in promotional pricing within the past fiscal year.
6. Change in the company's website hits due to the most recent advertising campaign.

Answers: Lead, Lead, Lag, Lead, Lag, Lag.

23

analysis designed to provide lead information. There are two fundamental ways that predictive analytics can predict the future and thus generate lead information. They are passive prediction and active prediction.

Passive Prediction Many attempts by analysts to predict the future are done as passive observers. An analyst may wish to predict whether a visitor to a website will make a purchase, based on observed browsing behavior within that website. In making her prediction, the analyst assesses the relationship between a purchase and observed information only; she does not consider how purchasing behavior might be affected if the visitor's browsing experience were altered or managed. We define predictive analytics of this type as passive prediction.

Passive prediction is the use of predictive analytics to make predictions based on actual and/or hypothetical data for which no variables are exogenously altered. That is, the data are passively observed. To round out this definition, note that a variable in a dataset is said to be **exogenously altered** if it changes due to factors outside the data-generating process that are independent of all other variables within the data-generating process. If

instead of passively observing the data, our analyst alters a visitor's browsing experience by changing a banner ad to an automatic pop-up ad, this change would constitute an exogenous alteration to that visitor's ad exposure.

passive prediction The use of predictive analytics to make predictions based on actual and/or hypothetical data for which no variables are exogenously altered.

exogenously altered A variable in a dataset that changes due to factors outside the data-generating process that are independent of all other variables within the data-generating process.

Using the definition for passive prediction, let's expand our website visitor example. The analyst may have data on visits by many individuals, and then use those data to establish a relationship between browsing behavior and purchasing (e.g., by estimating a regression model). Then, using that established relationship, she can make predictions for a given visitor based on his observed browsing behavior, and she can make predictions for a generic visitor based on hypothetical browsing behavior (i.e., what if a visitor browses page 2 for 30 seconds, page 5 for 45 seconds, etc.?). This use of predictive analytics is passive prediction.

Pattern discovery (also known as data mining for large data), when used to make predictions, is generally used for passive prediction. Conceptually, if you discover a distinctive pattern among a set of variables in a given dataset, it is natural to expect that this pattern will emerge again when the same variables are collected without interference. For example, using data on high school student demographics and graduation outcomes, you may discover a pattern in which the level of education attained by a student's mother relates to whether that student graduates from high school.

Recognizing this pattern, you may build a model for classification, in which students are classified as "graduate" or "not graduate" based on the level of their mother's education (e.g., those with mother's education more than 12 years may be labeled "graduate," and "not graduate" otherwise). Of

course, it is unlikely that a classification model as simple as this one will be sufficient to make good predictions. However, the key take-away from this example is that if such a classification model were used to make predictions, it would do so in a passive way. You would observe or hypothesize the level of education for a student's mother, and then use the model to make a prediction about whether that student will graduate. At no point in this process would you consider exogenously changing the level of education the mother attained, and then predicting the effect of such a change on graduation.

A highly visible example of passive prediction is weather forecasting. To predict the weather, forecasters use predictive analytics models and techniques that

24

discover and utilize patterns in large weather datasets. In forming predictions, the forecasters take a passive approach, making predictions based on the current (and past) weather they observe. Weather forecasters do not make predictions for, say, tomorrow's temperature if a massive inferno breaks out on the west end of the city tonight. This is because the instance of the inferno is an exogenous alteration to, say, ground conditions, since it comes from outside the general data-generating process for the weather.

Passive prediction is done using a wide range of predictive analytics models designed for pattern discovery and data mining. Commonly used are neural networks and decision trees, as well as regression. While we will not discuss such models in detail in this book (except for regression), the basis on which analysts generally choose among competing models for passive prediction is *model fit*. They seek the model that can most closely match the data according to some metric. This criterion is sensible given the goal of passive prediction; the analyst simply wants as accurate a prediction as possible following a given, unaltered, data-generating process.

Passive prediction has many applications in business, ranging from predicting customers most likely to drop service to employees most likely to be successful in the company. A famous example of passive prediction utilized customer data on purchases. Analysts at Target used credit card

purchasing data to make passive predictions about whether a woman was pregnant. In one case, a father visited Target to yell at its personnel for sending his daughter coupons for baby-related items. He later learned that his daughter was, in fact, pregnant but had not yet told him. He subsequently apologized to Target.

Active Prediction Often in business and elsewhere, we want to predict the consequences of some action on an outcome of interest. A student may want to predict the impact on her final grade if she increases her study time by one hour per week. Here, the action is the increase in hours studying, and the outcome of interest is her final grade. This type of prediction is active prediction. **Active prediction** is the use of predictive analytics to make predictions based on actual and/or hypothetical data for which one or more variables experience an exogenous alteration. For the grade example, the data-generating process has the final grade depending on study hours and other variables; the prediction we want involves exogenously altering the process by changing the number of hours studied, independent of all other variables influencing the final grade.

active prediction The use of predictive analytics to make predictions based on actual and/or hypothetical data for which one or more variables experience an exogenous alteration.

Making active predictions requires establishing a causal relationship between variables. If variable X changes exogenously and this change affects our prediction for variable Y , this impact must be due to a causal relationship between the two variables. There is no movement in any other variable that could explain why the prediction changes with X . Hence, we must know the causal relationship between Y and X in order to make active predictions for Y based on exogenous changes in X .

For example, suppose the outcome you care about is your body weight two months from now; call this variable Y . Currently, you do not eat whole grains, but are considering switching to a whole-grain-only diet; call this change in diet X . Then, you may want to make an active prediction of X on Y .

To do this properly, you must understand the causal impact of changing to a whole-grain only-diet on future body weight.

We conclude this section by providing practical examples of passive and active prediction in [Communicating Data 1.5](#).

25

COMMUNICATING DATA 1.5

PASSIVE AND ACTIVE PREDICTION IN POLITICS AND RETAIL

One of the most prominent uses of *passive* prediction is in political races. Gallup and many other political consulting businesses spend millions of dollars every year to predict the outcomes of elections.

One way they can do this is through passive predictions. For example, we can divide people into different subsets of the population based on their personal characteristics: gender, union participation, race, income level, etc. Since we can observe only an individual's personal traits (we cannot change a person's race, for example), we use passive prediction to make inferences about which groups might be more likely to vote for each political party.

To illustrate, in 2009 Gallup conducted a study about the effect of gender on political preferences. The pollsters found that in the United States, women are more likely than men to favor Democratic candidates, regardless of age, race, and marital status. Using these data, we could predict that a group of women would be more likely, relative to men, to favor a Democrat in an upcoming election, and could plan a political strategy around that information.

An interesting business application of *active* prediction involves the background music in retail stores. You might not notice music playing when you go shopping, but studies show that it can significantly influence the sales. In a study published in the *Journal of Marketing*, researchers actively varied music tempo played in stores from very slow to quick, and observed the results on shopping behavior and sales. The study indicated that quick-tempo music influenced store patrons to move faster through the store; slower music caused

patrons to shop more slowly while in the store. The researchers even found that playing slow music caused shoppers to move more slowly than having no music at all. As might be expected, the researchers also found that slower music resulted in higher sales, while fast music resulted in lower sales. These findings suggested that a retail store owner may be able to increase sales by strategically altering what's playing in the background.

Active Prediction for Business Strategy Formation

LO 1.8 Recognize questions pertaining to business strategy that may utilize (active) predictive analytics.

As noted at the start of the chapter, predictive analytics is highly useful in business strategy formation: The decision maker can forecast the outcomes of alternative strategies and then choose the strategy that is best according to his objectives. We can clarify this idea further now that we have distinguished between passive and active prediction.

Predicting an outcome for alternative strategies requires the application of active prediction. This is because a shift in strategy by a firm is essentially an exogenous shock to the data-generating process that was determining the outcome and associated variables. To accurately predict an outcome(s) for a range of competing strategies, you must be able to establish the causal effect of those strategies on that outcome.

To illustrate how predicting the effects of business strategies relies on active prediction, and ultimately causal inference, let's consider a pricing problem. Any firm selling a product or service must make a strategic decision about the price it charges; it often will revisit this decision throughout its existence. Suppose an online firm is currently selling its product for \$20. Over the past year, it has varied its price between \$10 and \$30 due to changes in demand conditions (e.g., holiday shopping) and supply conditions

(changes in supplier prices). Under the current demand and supply conditions, the firm is considering

26

raising its price from \$20 to \$25. However, before doing so, it wants to predict what this price hike would do to profits.

Making this prediction requires the firm to understand cause and effect for price. The \$5 price hike is the cause, and the change in profits is the effect. To determine this relationship and ultimately make a prediction about the effect of the price hike, it can be tempting to simply measure how profits moved with price, using data on prices and profits. In fact, there are still many instances in which firms and researchers make this type of mistake even to this day. However, the leap from correlation to causality is a large one, and can lead to grossly incorrect predictions.

Suppose demand for the firm's product is particularly high during the holidays, and price was set at \$15 during that time, motivated by the belief that shoppers would be particularly active and price sensitive during that time. Suppose also that demand for the product is low in the summer, and price was set at \$25 during that time, motivated by the belief that those who were interested in the product then were not aggressively comparing prices. Under this scenario, the data are likely to show high profits when price was \$15 and low profits when price was \$25. It is tempting to conclude, then, that raising price lowers profits.

However, is such a conclusion valid for a given set of market circumstances? For example, does this conclusion imply that raising price during the holidays would reduce profits? Or more generally, does it imply that raising price from \$20 to \$25 at a given point in time will be detrimental to profits? The answer to these questions is "not necessarily." Understanding why and knowing the reasoning and analysis required in these types of predictions is where we will focus our attention for the remainder of the book.

RISING TO THE **data**CHALLENGE

Navigating a Data Dump

Let's return to the Data Challenge posed at the start of the chapter: navigating a data dump. To begin this task, you should ascertain two key data features: the units of observation and the data-generating process. Different variables from different sources are often collected at different rates and levels. That is, their units of observation are different. For example, the Number of Likes may be a time series with the unit of observation being a week. In contrast, Product Sales may be a panel with the unit of observation being a store-month. Understanding the units of observation is crucial when trying to combine these variables into one dataset.

Here, you can add up the weekly “Likes” data for each month, and keep this figure constant across all stores (since it doesn’t vary by store). Combined data may look as follows:

MONTH	STORE	FACEBOOK LIKES	PRODUCT SALES
1	1	2347	28
1	2	2347	19
2	1	2782	33

27

In establishing the data-generating process, you should begin by thinking about all the variables that might influence Product Sales. The first variable you might want to include is Facebook “Likes,” as this is the variable whose effect you want to check. The next variable probably is price, since there is a clear economic relationship between price and sales. What other variables might impact sales? For example, do you think production costs are part of the data-generating process for sales? It may seem like a relevant variable since higher costs can lead to higher prices and lower sales; however, this means that if we know the price, changes in production costs don't have an impact on sales.

Taking time to think about the components of the data-generating process

becomes especially valuable as the number of variables in your dataset becomes very large. It can be very costly and wasteful to analyze relationships across hundreds, thousands, or even millions of variables when some simple reasoning can reduce the analysis to something much smaller and simpler.

SUMMARY

This chapter introduced the relationships among data, predictive analytics, and business strategy. It discussed data features by distinguishing structured and unstructured data, and by explaining unit of observation and the data-generating process behind the data that are observed. We then explained how data analysis is used for business, classifying this use into three broad categories: queries, pattern discovery, and causal inference. We discussed how data can provide lag and lead information, and distinguished between predictive analytics that provides passive and active predictions.

A key purpose of this chapter is to provide a strong perspective on the applications of data analysis within the very broad category of Business Analytics, and within the narrower category of Predictive Analytics. Understanding the difference between active and passive prediction is crucial to the remainder of the book, as it sets the foundation of the book's focus and highlights how the application of predictive analytics for business strategy formation requires the use of active prediction. The remainder of the book will explore how to conceptualize, execute, and communicate these types of predictions.

KEY TERMS AND CONCEPTS

[active prediction](#)

[association analysis](#)

[business analytics](#)

[business strategy](#)

[causal inference](#)

cluster analysis
cross-sectional data
dashboard
data
database
data-generating process
data mining
descriptive statistics
direct causal relationship
exogenously altered
indirect causal relationship
key performance indicator
lag information
lead information
outlier detection
panel data
passive prediction
pattern
pattern discovery
pivot table
pooled cross-sectional data
predictive analytics
query
report
scorecard
structured data

time series data

unit of observation

unstructured data

28

CONCEPTUAL QUESTIONS

1. Classify the following datasets as structured or unstructured. (LO2)
 - a. Jim: Age(22), Height(70in), Location(USA); Ann: Age(32), Height(65in), Location(Netherlands); George: Age(47), Height(73in), Location(UK); Gloria: Age(61), Height(68in), Location(Switzerland)
 - b. Texts: "It's cold today" "Hello?" "Where are you?"; Date: 1/14/15, 4/17/15, 9/25/15; Price: \$0.10, \$0.22, \$0.28
 - c.

COLOR							
	RED		GREEN		VIOLET		BLUE
Length							
	15in		17in		22in		14in
Weight							
	5lbs		7lbs		2lbs		4lbs
Width							
	10in		12in		8in		12in

d.

SALES	TIME		SALES		PRICE		SALES
1440	16:45		1992		\$8.42		1924
Time	Price		Store		Sales		Time
14:17	\$4.45		22		1312		10:55
Price	Store		Time		Store		Price

\$9.99	18		11:14		9		\$11.14
Store	Time		Sales		Price		Store
16	12:12		2122		\$7.18		7

2. For the following datasets, determine the unit of observation and the data type (cross-sectional, pooled cross-sectional, time series, or panel). (LO3)
- a.

YEAR	STORE	PROFIT	MANAGER	CUSTOMERS
2015	1	\$22,810	Anderson	411
2015	2	\$32,416	Hietpas	593
2016	1	\$25,983	Anderson	460
2016	2	\$19,542	Wozniak	384

29

b.

MONTH	PC SALES	TABLET SALES	ACCESSORIES SALES	WEBSITE HITS
1	\$3.4 million	\$4.6 million	\$1.3 million	1.2 million
2	\$3.7 million	\$4.9 million	\$1.5 million	1.3 million
3	\$3.2 million	\$5.0 million	\$1.6 million	1.3 million
4	\$3.5 million	\$4.8 million	\$1.4 million	1.2 million

c.

YEAR	NAME	SALARY	TENURE	AGE
2015	Herbert Grant	\$48,000	?4 years	32
2015	Dianne Lawson	\$53,000	?7 years	39
2016	Michael Daniels	\$67,000	10 years	45
2016	Kelly Harper	\$72,000	8 years	34

d.

--	--	--	--	--

FACTORY	UNITS PRODUCED	PRODUCTION COSTS	NUMBER OF EMPLOYEES	YEAR BUILT
1	832	\$12.3 million	210	1993
2	1,105	\$15.2 million	315	1998
3	512	?\$7.2 million	141	2005
4	916	\$11.4 million	253	1987

3. Characterize the following data analyses as a query, pattern discovery, or causal inference. (LO5)
- The correlation between sales and price is 0.21.
 - The rate at which men aged 20–29 made a purchase is 0.04.
 - A 10% increase in prime time advertising will increase sales by 3%.
 - Sonya Barber and Wesley Santiago both have incomes that are at least \$1,000,000 more than anyone else's.
 - Moving across the country before age 30 is associated with having a college education.
4. Explain the difference between lead and lag information. (LO6)
5. Explain the difference between passive and active prediction. (LO7)
6. Characterize the following predictions as passive or active prediction. (LO7)
- Tom predicts his body weight next month after changing to a whole-grains-only diet this month.
 - Ann predicts sales for her product next week based on the number of visits to her website in the past five days.
 - Laura predicts Twitter traffic for her company over the next 10 days after launching a new advertising campaign last week.
 - Alex predicts total revenues at Macy's next month using information on Macy's credit card purchasing last week.
 - John predicts the winner of a local election based on answers to a recent survey given by local voters.

-
7. Why is the application of predictive analytics an ideal complement to the formulation of business strategy? (LO1)

8. A manager claims that increases in advertising expenditure will surely raise the firm's profits, citing his sense that people find the firm's ads entertaining. (*LO1*)

 - a. Sketch how you might refute this claim using:

 - i. A theoretical argument
 - ii. Data
 - b. Why might the refutation using data be more convincing?
9. List three factors that are likely part of the data-generating process for the number of years an employee works at a firm. (*LO4*)
10. A grocery store manager is interested in the data-generating process for her store's weekly soda sales. She believes factors impacting these sales include price, product placement, and whether the week contains a holiday. Write out a formal representation of the data-generation process for weekly soda sales that incorporates these and additional factors. (*LO4*)
11. Which of the following business questions requires the use of active prediction? (*LO8*)

 - a. Are older people more or less likely to buy our product?
 - b. Will a new celebrity endorsement enhance sales?
 - c. How will profits respond to a change in product placement in the store?
12. As a manager at a major insurance company, Meredith asks two of her analysts to help answer two separate questions regarding a recent car insurance client. She asks the first analyst, Darryl, to predict using demographic characteristics (e.g., the client's age, education, etc.) the likelihood that the client will be in an accident in the next five years. She asks the second analyst, Amanda, to predict the effect of a 10% cut in the client's premium on the likelihood that the client switches providers in the next three years. Which analyst is making an active prediction? (*LO8*)

QUANTITATIVE PROBLEMS connect

13. Properly load the following data into an Excel spreadsheet. Clearly explain what are (1) the unit of observation and (2) the data type. (*LO2, LO3*)

Score(47), Year(2015), Vote(Yes), Name(Laurene Horton); Score(83),

Year(2015), Vote(No), Name(Wilson Zimmerman); Score(35), Year(2016), Vote(No), Name(Jeffrey Wade); Score(48), Year(2016), Vote(No), Name(Kayla Snyder); Score(26), Year(2015), Vote(No), Name(Jeffrey Wade); Score(91), Year(2015), Vote(Yes), Name(Candice Graves); Score(62), Year(2016), Vote(Yes), Name(Candice Graves); Score(52), Year(2015), Vote(Yes), Name(Kayla Snyder); Score(83), Year(2016), Vote(No), Name(Laurene Horton); Score(76), Year(2016), Vote(No), Name(Wilson Zimmerman)

- 14.** Construct a scorecard for the dataset *Scorecard.xlsx* to assess performance on the following objectives. (LO6)
- Average regional revenue is at least \$200,000.
 - Average regional growth is at least 5%.
 - Returns are no more than \$10,000 in any store.

Dataset available at www.mhhe.com/prince1e

31

-
- 15.** Access the dataset *Sales and Costs.xlsx* and answer the following questions. (LO5)
- Calculate these descriptive statistics.
 - Mean of sales
 - Variance of materials costs
 - Covariance of labor costs and materials costs
 - Mean of labor costs
 - Total sales
 - Calculate at least two more descriptive statistics for this dataset.

Dataset available at www.mhhe.com/prince1e

- 16.** For the dataset *Sales and Costs.xlsx* answer the following queries. (LO5)
- What are average sales when labor costs exceed \$50,000?
 - What are minimum sales when materials costs are less than \$20,000?
 - What region had the highest labor costs?
 - What region had the lowest sales?

e. What was the greatest difference in sales across any two regions?

Dataset available at www.mhhe.com/prince1e

17. Using the dataset *Lead and Lag.xlsx*, answer the following questions and explain whether the answers can be categorized as lag or lead information. (LO6)
- Which customers are most likely to drop your service?
 - Which region had the most customers last year?

Dataset available at www.mhhe.com/prince1e

¹ In business, some will draw distinctions between strategies and tactics, claiming that a pricing decision is typically a tactic rather than a strategy. We simplify the discussion by labeling what some may call tactics as “simple” strategies.

Reasoning with Data

LEARNING OBJECTIVES

After completing this chapter, you will be able to:

- LO2.1 Define reasoning.**
- LO2.2 Execute deductive reasoning.**
- LO2.3 Explain an empirically testable conclusion.**
- LO2.4 Execute inductive reasoning.**
- LO2.5 Differentiate between deductive and inductive reasoning.**
- LO2.6 Explain how inductive reasoning can be used to evaluate an assumption.**
- LO2.7 Describe selection bias in inductive reasoning.**

Chapter opener image credit: ©naqiewei/Getty Image

dataCHALLENGE Testing for Sex Imbalance

You have just begun work as an analyst at an online publishing firm, and it has just put up for sale a new “Do It Yourself” (DIY) book. The author intends to write several follow-up books, but hopes to learn about the audience of his first book as he begins to design the next. After reading the book, one of your colleagues at the firm claims the book is heavily tailored to men, and consequently the rate of purchase among women viewing the book on the website will be very low, around 3%. You are skeptical of this claim, but rather than make a theoretical rebuttal, you believe data are your best bet to disprove it. Your publishing firm collects data on the sex of its visitors and whether they purchased, and the current data sample has the purchase decisions of 500 female visitors who viewed the book on the website. Outline the reasoning process you might follow that utilizes this data sample to test your colleague's claim.

33

Introduction

This chapter builds the foundation for the use of reasoning when drawing conclusions from data analysis. The purpose of the chapter is to build a general framework for how one should think about data analysis. By using a framework centered on reason, we can effectively answer questions like:

- “What must I believe in order to draw a general conclusion about what is going on with my company, based on what I see in a dataset?”
- “How confident am I in the conclusions I have just drawn based on the data I have seen?”
- “What is the line of reasoning that took us from that number to that conclusion?”

To build this reasoning “thought structure,” we will define reasoning and discuss in detail two key types of reasoning: deductive reasoning and inductive reasoning. We will then show how both types of reasoning play a crucial role when we try to draw general conclusions based on the data we observed. We will show how deductive reasoning can lead to empirically testable conclusions,

and how we use data and inductive reasoning to test these conclusions. Throughout this chapter and the remainder of the book, we will be incorporating *Reasoning Boxes*. These summarize the reasoning behind some of the main concepts in the book.

Some of the concepts we present in this chapter can be a bit abstract, but they all ultimately are highly valuable for business application. To help make the upcoming concepts more concrete, we start with a simple example. Consider a coin, say, a quarter. Any quarter has one side we call “heads” with the other side “tails.” Now, consider a standard assumption we might make about our quarter: “The quarter is fair.” We assume that, upon flipping the quarter in the air, the probability it lands with heads facing up is 50%, and the probability it lands with tails facing up is 50%. However, should we take our assumption that the coin is fair on faith? How could we test this assumption? One way is with data—we flip the coin. If we flip the coin five times and see heads each time, would we doubt the coin is fair? Intuitively, believing the coin is fair leads us to expect certain outcomes when we start flipping the coin; we expect a “balanced” number of heads and tails. We flip the coin and see whether or not our expectations are met, and then decide whether or not we believe, after flipping the coin several times, the coin is fair.

We can apply all the reasoning in this chapter to this coin example. What remains is to clearly define and describe each component of reasoning being used. By doing so, we can explain and debate how we arrive at data-driven conclusions.

What is Reasoning?

LO 2.1 Define reasoning.

When confronted with a new dataset, it is easy and natural to begin analysis by calculating classic summary statistics, such as means and standard

deviations. Interpreting these numbers is straightforward; they correspond to clear definitions. For example, if you determine

34

the mean of X in your sample is 20, this literally means that the sum of your observed values for X divided by the number of times you observed X is 20. Mathematically, we write this as:

$$\bar{X} = \left(\frac{1}{N}\right) \sum_{i=1}^N X_i = 20$$

If you accept this definition, then arriving at the conclusion that the mean is 20 is nothing more than a simple exercise in algebra, typically done by your computer or calculator.

Data analysis, though, involves more than just calculations using mathematical and statistical definitions. It also can involve more powerful statements and conclusions rooted in the statistics that are generated. Relevant examples for this book are statements about causality—for example, an increase of \$1 million in R&D will generate three new patents. The three-to-one relationship comes from mathematical and statistical calculations (e.g., through a regression, detailed in [Chapter 5](#)); however, their interpretation as quantifying a causal effect between R&D and patents requires more than just calculations. Such an interpretation requires the use of reasoning.

“Reasoning” is a term with which everyone is familiar. We could argue that humans’ advanced ability to reason is the key feature that separates us from every other creature on Earth. Despite this, many of us have only a vague notion as to what constitutes reasoning, and one can find examples of flawed reasoning almost anywhere. (The Internet is a good place to start.)

Before developing a more formal understanding of reasoning, consider the following two arguments, each relying on your ability to reason in order to convince you they are correct:

1. The company's profits are up more than 10% over the past year. An increase in profits of 10% is the result of excellent management. You

were the manager over the past year. Therefore, I conclude that you engaged in excellent management last year.

2. Ten of your 300 employees came to me with complaints about your management. They indicated that you treated them unfairly by not giving them a raise they deserved. Therefore, I conclude that all of your employees are disgruntled with your management.

In presenting those two examples, the goal is not for you to make a definitive decision about which you believe (if either). Rather, the goal is to begin thinking about, and distinguishing, different “lines” of reasoning. By the end of this section, you will be able to clearly distinguish the types of reasoning used in these two examples. You will then be able to carefully establish *why* you either believe or question the claims they are making.

We define **reasoning** as the process of forming conclusions, judgments, or inferences from facts or premises. Reasoning and logic are often used interchangeably, and the difference between the two is quite subtle. If a distinction is to be made, reasoning is a thought process, while **logic** is a description of the rules and/or steps behind the reasoning process. Given the similarity in their meaning and application, we will follow the common practice of using the terms “reasoning” and “logic” to mean the same thing in this book.

reasoning The process of forming conclusions, judgments, or inferences from facts or premises.

logic A description of the rules and/or steps behind the reasoning process.

The study and use of reasoning have a rich history in philosophy, dating all the way back to Aristotle. While some people view reasoning as the only way to process information

and make decisions, others take alternative approaches, such as following a hunch or “going with their gut.” The business environment is populated with

both types of people, at all organizational levels. Nevertheless, when people seek to communicate information and conclusions, doing so via reasoning has a much greater potential for impact and consensus.

In the context of data analysis, use of reasoning allows all those processing the information to clearly see how you can go from a dataset to a meaningful conclusion, rather than rely on one or more people's instincts or feelings. Consequently, especially for data analysis, it is crucial to have a basic understanding of the fundamental components of reasoning and be able to apply them when producing and assessing data-driven conclusions.

Deductive Reasoning

Broadly speaking, reasoning is divided into two major types, and all of us engage in both types. The two types of reasoning are *deductive reasoning* and *inductive reasoning*. They are fundamentally different, but both play an important role when properly interpreting, and drawing conclusions from, data analysis.

DEFINITION AND EXAMPLES

LO 2.2 Execute deductive reasoning.

Deductive reasoning goes from the general to the specific. It is also known as *top-down logic*. Deductive reasoning seeks to prove statements of the form “If A, then B.” For example, you may use deductive reasoning to prove the statement: “If our customers are very sensitive to price, then a price increase will ultimately lower our stock price.”

deductive reasoning Reasoning that goes from the general to the specific; also known as *top-down logic*.

Such reasoning always implies three underlying components: assumptions (“If A”), methods of proof (“then”), and conclusions (“B”). The

process conceptually is straightforward, as illustrated in [Figure 2.1](#).

After understanding the structure of deductive reasoning, the next task is to understand how to apply it in practice. That is, in business, how do you properly support your conclusions using deductive reasoning? The purest applications of deductive reasoning are in the field of mathematics; a basic understanding of how it is used there can provide a strong foundation for applying it in business. Two of the most utilized approaches for proving “If A , then B ” in mathematics are *direct proofs* and *transposition*. These methods are also particularly useful in business due to their intuitive appeal.

A **direct proof** literally walks across the flow chart in [Figure 2.1](#), filling in all the details along the way. To see how this works, consider a simple mathematics application. Let's prove the following statement: “If X and Y are odd numbers, then their sum ($X + Y$) is an even number.”

direct proof Proof that begins with assumptions, explains methods of proof, and states the conclusion(s).

This statement may seem intuitively obvious; you can pick any two odd numbers (e.g., $X = 5$ and $Y = 9$), add them up, and see the sum is always an even number ($X + Y = 14$, which is even).

FIGURE 2.1 The Deductive Reasoning Process



36

However, failing to find a contradiction is not the same as proving a statement is generally true. How then can we use a direct proof to prove this statement? Consider the following line of reasoning:

1. X and Y are odd numbers.
2. If X is an odd number, we can always write it as $X = 2 \times K + 1$, where K is an integer. For example, if $X = 13$, $X = 2 \times 6 + 1$.

3. If Y is an odd number, we can always write it as $Y = 2 \times M + 1$, where M is an integer.
4. When we add X and Y , we get: $X + Y = 2 \times K + 1 + 2 \times M + 1 = 2 \times K + 2 \times M + 2 = 2 \times (K + M + 1)$.
5. $K + M + 1$ is an integer, so this means $X + Y$ is 2 times an integer.
6. Any number that is 2 times an integer is divisible by 2.
7. This means $X + Y$ is even.

The above line of reasoning is an acceptable formal proof of the originating statement that uses a direct proof. The added details along the way include definitions (of even and odd numbers, in steps 2, 3, and 6), simple algebra (step 4), and a property of integers (a sum of integers is an integer, in step 5).

As we step outside mathematics, the methods of proof can become less formal. Rather than relying on formal definitions or rigid mathematical rules, they may rely on generally accepted principles or even common sense. Consider a recent decision by McDonald's to offer breakfast all day. The underlying deductive argument for this decision may very well have been: "If our stores offer breakfast all day, revenues will increase." A direct proof of such a claim may look as follows:

1. McDonald's stores offer breakfast all day.
2. The addition of breakfast during lunch/dinner hours implies more choices for customers wanting to eat out during lunch/dinner hours without losing any choices previously available.
3. Consumers already choosing McDonald's during lunch/dinner hours can continue to buy the same meals.
4. Consumers not choosing McDonald's during lunch/dinner hours may start eating at McDonald's with the new food options.
5. By retaining current consumers and adding some new consumers, McDonald's sales will increase overall.

Here, step 2 details what the assumption implies for a given store's menu. Steps 3 and 4 appeal to common economic sense, and step 5 implicitly

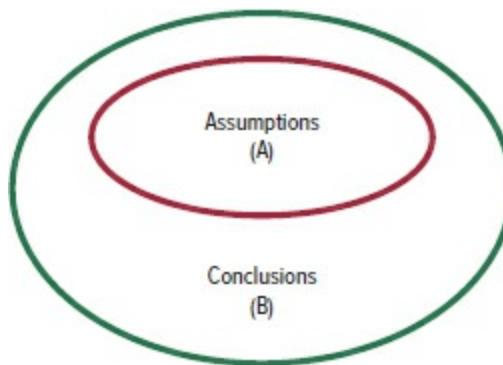
applies basic algebra.

Our McDonald's proof is well outside the realm of mathematics, and so expectedly suffers from a lack of rigor. In fact, you may disagree with the line of reasoning presented (and for good reason). We will revisit this proof below.

While direct proofs often are sufficient to prove a point logically, it can sometimes be easier and/or more effective to take an alternative approach. A popular alternative is to use transposition. Simply defined, **transposition** is the equivalence between the statements "If A , then B " and "If not B , then not A ." In other words, it means that any time a group of assumptions implies a conclusion (A implies B), then it is also true

transposition Any time a group of assumptions implies a conclusion (A implies B), then it is also true that any time the conclusion does not hold (not B), at least one of the assumptions must not hold (not A).

FIGURE 2.2 Illustration of "If A , then B ," and equivalently, "If not B , then not A "



that any time the conclusion does not hold (not B), at least one of the assumptions must not hold (not A).

Perhaps the most convincing way to illustrate this idea is through a picture. In [Figure 2.2](#), there are two ovals, A and B , with A inside of B . One

way to verbalize what this picture means is to say that “Anything inside of A is also inside of B ,” or in short, “If A , then B .” However, notice we could also verbalize what this picture means by saying “Anything outside of B is also outside of A ,” or in short, “If not B , then not A .”

As we did with a direct proof, we can first see how a proof via transposition works through a mathematics application. Let's now prove the following statement via transposition: “If X^2 is even, then X is even.” Again, this statement may seem obvious after just checking a few numbers. For example, 16 is even and equals 4 squared, and 4 is in fact an even number. How can we prove this statement generally using transposition? Consider the following line of reasoning:

1. Suppose X is not an even number; in that case, it is instead an odd number.
2. If X is an odd number, we can always write it as $X = 2 \times K + 1$, where K is an integer.
3. $X^2 = (2 \times K + 1)^2 = 4K^2 + 4K + 1$.
4. $4K^2 + 4K = 4(K^2 + K)$, and so is divisible by 2.
5. $4K^2 + 4K$ is an even number.
6. $X^2 = 4K^2 + 4K + 1$ is an even number plus 1, meaning it is an odd number.

The above line of reasoning is an acceptable formal proof of the originating statement that uses transposition. Here, we start with the opposite of the conclusion, and show that it leads to the opposite of the assumption. The added details along the way include definitions (of even and odd numbers, in steps 2, 5, and 6) and simple algebra (steps 3 and 4).

We can again step outside mathematics to see how we can use transposition to prove simple “If A , then B ” statements. The intuitive appeal of using transposition rather than direct proof lies in the fact that it sometimes is easier to prove a point by showing that disbelief of a conclusion requires clear disbelief of an assumption, rather than walking down a direct line of reasoning. Transposition can be particularly effective if an assumption seems

indisputably obvious; it then becomes a powerful affirmation of the conclusion to show that, were the conclusion not true, a seemingly obvious assumption must not be true. Transposition also can be useful in properly assessing the validity of a statement, since it invites consideration of other, unstated assumptions that may also need to be true in order for the conclusion to hold.

38

To see these points, consider again the claim for McDonald's: "If our stores offer breakfast all day, revenues will increase." A proof via transposition for the above statement may look as follows:

1. McDonald's stores revenues will not increase.
2. This means total revenues from current and new consumers will not increase.
3. This means either there will be no new consumers or revenue from current consumers will decrease.
4. This means there could not have been an expansion in the menu.
5. McDonald's stores do not offer breakfast all day.

In the above transpositional proof, there are two key take-aways. First, it takes the approach of showing that, if the conclusion does not hold, the assumption clearly cannot either. In short: "If McDonald's revenues do not increase, then the company could not have been offering breakfast all day." As another example, a teacher may claim "If you adequately study for the test, you will pass." Then, she may casually "prove" this statement by saying "If you fail the test, it is obvious you did not adequately study for it."

The second take-away from the above proof is that the presence of unstated assumptions can be clearer in this form. Following the transposition highlights the loose link between menu expansion and revenues from new and current consumers, most glaringly the latter. Is it a certainty that people will continue to buy the same meals when the menu expands? McDonald's should not lose these customers, since they can still buy the meals they bought before; but they may choose to buy some of the new breakfast options

instead of the traditional lunch/dinner meals. And if the breakfast options are less expensive, revenues could go down. Hence, the implicit assumption is that current consumers will not switch to the breakfast offerings. At least early into this strategy shift by McDonald's, franchisees were finding this implicit assumption to be false, resulting in declining revenue.

If an implicit assumption is discovered, and it is possible it may not always hold, it is advisable to then add it to the original statement. For the McDonald's example, the statement should then read: "If our stores offer breakfast all day and *current lunch/dinner consumers do not switch to breakfast options*, revenues will increase."

We summarize direct proofs and transposition in [Reasoning Box 2.1](#).

REASONING BOX 2.1

DIRECT PROOF AND TRANSPOSITION

Two key methods for making deductive arguments, direct proof and transposition, are as follows.

Direct proof:

1. State assumptions
2. Explain methods of proof: These include definitions, formulas, generally accepted principles, common sense, etc.
3. State conclusions

Transposition:

1. Assume the opposite of the conclusion (*not B*)
2. Explain methods of proof: These include definitions, formulas, generally accepted principles, common sense, etc.
3. State assumption(s) that is (are) violated (*not A*)

COMMUNICATING DATA 2.1

DEDUCING GUILT AND INNOCENCE

One of the most prevalent uses of deductive reasoning is in the application of law. Lawyers representing a defendant will often call upon a fundamental deductive argument to argue their client “didn’t do it.” The argument is: “If the person committed the crime, then he or she must have been present at the crime scene when the crime was committed.” Of course, such a statement applies only to physical crimes (murder, robbery), and not necessarily other types of crimes (identity theft). The transposition of this statement is familiar to many, and can be stated as, “If the person was not present at the crime scene when the crime was committed, then he or she must not have committed the crime.” This is the basis of an alibi, where a defendant claims to be elsewhere when the crime occurred in order to claim innocence. If the defense can provide sufficient evidence that the defendant was elsewhere (if it can provide a convincing alibi), then using simple deductive reasoning, the defendant has an extremely strong case for innocence. [Figure 2.3](#) illustrates this concept.

FIGURE 2.3 Visualization of the Alibi



2.1

Demonstration Problem

Consider the following claim “Firms that consistently attract top talent will inevitably be profitable in the long run.” Restate this in the form of a deductive argument. Then, show how a proof by transposition will likely fail due to at least one missing assumption.

Answers:

Deductive argument: If a firm consistently attracts top talent, then it will be profitable in the long run.

Transposition: If a firm is not profitable in the long run, it does not consistently attract top talent.

Missing assumption (example): The firm must also retain top talent. Thus, a firm may be unprofitable in the long run and consistently attract top talent, because it is not able to keep that talent.

Regardless of the approach used to prove a deductive statement, the appeal of using deductive reasoning to make an argument is in its concreteness. If there is disagreement with a conclusion, then there are only two possible sources: the method(s) of proof or the assumption(s). Consequently, a deductive argument that clearly lays out the assumptions, methods of proof, and conclusions allows for clear lines of assent and disagreement for those assessing the argument. To further see its value in a business environment, consider two different ways of arguing in favor of sales growth over the next year:

Method 1: “Suppose we continue investing in advertising online to the 18–34 demographic, and we see just a slight uptick in the effectiveness of our advertising to that group. Further, assume there are no notable drop-offs among other demographic groups. This means we would have stable sales in

some areas and improved sales in others, and consequently, I expect sales to grow over the next year.”

Method 2: “The young demographic loves our product, and we have been pushing our online advertising toward that group. My daughter is in that group, and she is excited about our upcoming product line; I’m excited about our upcoming product line! I firmly believe sales will grow over the next year.”

Although there is always a place in business for inspirational speeches (Method 2), they are not well suited for supporting a stated forecast. Within Method 1, we can see a clear set of assumptions, and the speaker is arguing that these assumptions lead to the conclusion. If an audience member believes sales will not grow over the next year, he must either disagree with the idea that the proposed assumptions lead to the conclusion or disagree with one of the assumptions. For example, he may argue that he does expect notable drop-offs among other demographic groups, and these drop-offs will ultimately hurt sales. For Method 2, there is no clear line of reasoning that leads to the conclusion. (Ask yourself: What are the key assumptions that lead to the sales forecast?)

We conclude this section on deductive reasoning by considering ways to resolve disagreements about a conclusion. As noted above, if deductive reasoning was used, such disagreements must be rooted in the method(s) of proof or the assumption(s). How are such disagreements settled? Often, methods of proof in business call upon obvious statements/definitions and generally accepted ideas (e.g., if price goes up, quantity sold will decline). Of course, this need not always be the case, and so disagreement about a definition or supposedly accepted idea may occur. However, disagreement about a method of proof is more likely to be due to an unstated but necessary assumption. As we did in the McDonald’s example, the way to resolve such a disagreement is to simply add this assumption in the claim (e.g., “If A_1 and A_2 , then B .”). Once this is done, the potential for disagreement moves from methods of proof to just the assumptions.

If disagreement about a conclusion stems from disagreement about one or more assumptions, how can this disagreement be settled? That is, how can we

evaluate assumptions that are disputable? Consider the claim “If our city's population increases by 2%, then sales will increase by at least 1%.” Suppose you agree with the methods of proof associated with that statement, but you do not believe the conclusion because you do not believe the assumption. You do not believe your city's population will increase by 2%. There are two main ways of resolving disputes about assumptions: (1) show robustness, and (2) assess consistency with a collected dataset.

Robustness in the context of a deductive argument is the persistent accuracy of a conclusion despite variation in the associated assumption(s). Put another way, a conclusion reached via deductive reasoning is robust if we are able to vary some or all of the assumptions that

robustness The persistent accuracy of a conclusion despite variation in the associated assumption(s) within the context of a deductive argument.

41

led to that conclusion and still reach the same conclusion. Perhaps you do not believe sales will increase by 1% because you believe your city's population will increase by only 1.2%, not 2%. Such a disagreement can be resolved if it can be shown that the conclusion is robust to variations in the assumption. For example, the conclusion may be true for population increases ranging from 0.7% and up. If so, we say the conclusion is robust to population increases all the way down to 0.7%.

Showing a conclusion is robust to a wide range of assumptions can be highly effective toward gaining consensus about its accuracy. However, many conclusions reached via deductive reasoning are not robust to certain variations in the associated assumptions. In such cases, one set of assumptions ($\{A_1, \dots, A_N\}$) might lead to one conclusion (C_1) while an alternative set of assumptions ($\{A'_1, \dots, A'_M\}$) might lead to a different conclusion (C_2). Here, the conclusion C_1 clearly is not robust at least to some alternative assumptions. How then do we decide whether to make assumptions $\{(A_1, \dots, A_N)\}$ instead of $(\{A'_1, \dots, A'_M\})$, and thus believe

conclusion C_1 instead of conclusion C_2 ? The answer is to assess each conclusion's consistency with a collected dataset, which we detail further next.

EMPIRICALLY TESTABLE CONCLUSIONS

LO 2.3 Explain an empirically testable conclusion.

Consider the following example. Suppose Chiquita Bananas is negotiating with a major grocery store chain, Kroger, about its product placement in the store. Chiquita's bananas are currently in a standard location among the produce, but the company is considering purchasing a premium location in the front of the store. Managers within Chiquita have two opposing points of view about such a move. Manager Group 1 believes that a premium location in the front of the store will make the average consumer more likely to take notice of Chiquita bananas, compared to their current location in produce. This group purports that the up-front location is sure to grab attention. Manager Group 2 believes that a premium location in the front of the store will make the average consumer less likely to take notice of Chiquita bananas, compared to their current location in produce. This group purports that most consumers aren't looking to make produce selections when they first walk into the store, and so will ignore a featured produce product; and when they do enter the produce section, they will not see Chiquita bananas.

Let's now map our Chiquita Bananas example into a basic deductive reasoning framework. Let's call Manager Group 1's belief that moving the product to the front of the store will make consumers take more notice Assumption 1 (A_1), and let's call Manager Group 2's belief that moving the product to the front of the store will make consumers take less notice Assumption 2 (A_2). Suppose also that both groups agree that greater notice will lead to greater sales. Then, we have two basic, competing arguments: "If A_1 , then C_1 " vs. "If A_2 , then C_2 ." Here, C_1 is the conclusion that sales will increase, and C_2 is the conclusion that sales will decrease. These alternative

conclusions (C_1 and C_2) are both examples of an empirically testable conclusion.

An **empirically testable conclusion** is a conclusion whose validity can be meaningfully tested using observable data. Conclusions C_1 and C_2 in our Chiquita Bananas example are both empirically testable. Suppose we have data on Chiquita Bananas sales in the current location. Then, suppose the company chooses to move their bananas to the new up-front location and collects sales data for that new location. With data on sales while the bananas are displayed in produce and in the front of the store, it should at least be conceptually clear that we can meaningfully test the validity of our competing conclusions. At first glance, we

empirically testable conclusion A conclusion whose validity can be meaningfully tested using observable data.

42

might argue that an increase in sales after the location change is consistent with conclusion C_1 being valid and conclusion C_2 being invalid. By making this distinction, we are effectively sorting between assumptions A_1 and A_2 . The invalidity of C_2 implies A_2 cannot be correct (using transposition: “Not C_2 implies Not A_2 ”). Hence, we would argue that A_2 is not consistent with the data. While this example is only conceptual (and quite informal), we formalize this process of testing conclusions and ultimately assessing assumptions later in this chapter, and in the chapters that follow.

Many conclusions pertaining to business outcomes are empirically testable. Examples include: (1) “If consumers are price sensitive, higher prices lead to lower revenues” and (2) “If consumers ignore all ads on web pages, changes in ad placement on a web page will not alter sales.” We can attempt to test the conclusions in both of these statements using data on prices and revenues, and data on ad placement and sales, respectively. In contrast, consider the statement: “If firms underestimate future profits, they will sometimes exit markets that would have proven highly profitable for them in the future.” Notice that the conclusion is difficult to empirically test;

to do so, we'd have to know how profitable a firm would have been had it not exited a market. We may learn something about this by looking at other firms; however, the only data that speak directly to this conclusion would be data on firm profitability after exiting, and such data cannot exist.

Thus far, our discussion of empirically testable conclusions has been rather abstract, relying on conceptual explanations of how testing occurs as opposed to statistical methods. Although we don't discuss statistical methods until the next chapter, we can begin building a more concrete foundation for the process of testing conclusions by focusing our attention on a particular subset of conclusions. Many of the practical conclusions we will test in this book concern *outcome probabilities*.

As illustration, consider again a quarter, and again suppose we assume “The quarter is fair.” The probability it lands with heads facing up is 50%, and the probability it lands with tails facing up is 50%. This assumption leads to many empirically testable conclusions. One is that, when flipping the quarter five times, “If the quarter is fair, the probability of seeing any number of heads (0–5) is as described in [Table 2.1](#). ”

How we might empirically test our conclusion is straightforward. We flip the quarter five times and record the heads that occur. The number of heads in our observed data serves as a test of our deductive conclusion (the probabilities in [Table 2.1](#)). If we observe a very high (5) or very low (0) number of heads in our five flips, we may question whether the probabilities in [Table 2.1](#) actually apply to the quarter we flipped, and thus whether our empirically testable conclusion is correct. If we decide our conclusion passes the test, then we have no reason to dispute the assumptions leading to that conclusion, but if we decide it does not pass, we must reconsider or reject at least some of the assumptions leading to that conclusion.

Making the actual decision about the validity of an empirically testable conclusion based on observable data is an application of *inductive reasoning*.

TABLE 2.1 Probabilities for Varying # of Heads When Flipping a Quarter Five Times

# OF HEADS	0	1	2	3	4	5
------------	---	---	---	---	---	---

Probability	3.125%	15.625%	31.25%	31.25%	15.625%	3.125%
-------------	--------	---------	--------	--------	---------	--------

Inductive Reasoning

LO 2.4 Execute inductive reasoning.

For the remainder of this chapter, we discuss the second major type of reasoning, inductive reasoning. Inductive reasoning plays an important role in distinguishing among competing sets of assumptions. We will explain inductive reasoning and its applications generally; how it can be used to sort through assumptions leading to empirically testable conclusions; and how it can be, and often is, improperly used.

DEFINITION AND EXAMPLES

LO 2.5 Differentiate between deductive and inductive reasoning.

Inductive reasoning is reasoning that goes from the specific to the general. It is also known as *bottom-up logic*. Broadly speaking, inductive reasoning is used to generalize based on what is observed. Most inductive arguments have the following basic form “Based on observing X, I believe that Y is generally true.”

inductive reasoning Reasoning that goes from the specific to the general; also known as *bottom-up logic*.

While other types of reasoning can be classified as inductive reasoning, the process of making a general statement based on specific observations is certainly the most well-known use of inductive reasoning. *It will be the version we use throughout this book.*

In contrast to deductive reasoning, inductive reasoning does not involve a proof. We cannot *prove* the general from the specific. We can only support or fail to support the general from the specific. Observing that 20 polar bears are white does not prove all polar bears are white; rather, it supports the claim that the entire polar bear population is white.

Inductive arguments stemming from business data analysis almost exclusively use information about a data sample to draw conclusions about a general population. The **population** is the entire set of potential observations (e.g., customers, sales in a month) about which we want to learn. A **data sample** is a subset of a population that is collected and observed. For example, a survey of 1,000 Americans that contains information on their age and income is a data sample of the population of all Americans. As a less obvious example, consider a data sample consisting of Apple's monthly iPhone sales. The corresponding population for this sample may seem less obvious since there isn't a larger set of monthly sales for iPhones from which we are sampling; there is one sales figure for each month, and we are observing it. However, we can conceptualize a larger set of *potential* monthly sales for iPhones (i.e., monthly sales that could have happened under different circumstances), and the sales we observed are drawn from this population.

population The entire set of potential observations about which we want to learn.

data sample A subset of a population that is collected and observed.

Inductive reasoning fits intuitively with data samples and populations. A data sample is the “specific,” which we observe, and the population is the “general,” about which we form conclusions. In business, it can be beneficial to understand key features concerning a wide range of populations, including current customers, potential customers, profits over time, sales regions, and so on. Businesses regularly collect data samples and apply inductive reasoning to draw conclusions about these populations.

Suppose a firm that conducts all its business over the Internet wants to know the average age of its customers. It would like to get this information with minimal intrusion on its customers' time. Consequently, over the period of a month, it sends one out of every twenty customers a brief questionnaire asking his/her age after a purchase is made. Let's assume this process results in 872 observations (purchasers whose age the firm observes), and the average age of these 872 people is 43.61. From here, the inductive reasoning is straightforward: "Based on observing an average age of 43.61 for my customers in my sample, I believe the average age of my customers in general is 43.61."

44

Inductive reasoning using data may seem simple. However, the given argument has an apparent weakness. Concluding that the sample average is the exact population average is a bold step; it invites the question: "How sure are you of that conclusion?" The answer requires knowledge of the degree of support for the conclusion. The **degree of support** (also called *inductive probability*) for an inductive argument is the degree of confidence in the conclusion resulting from the stated observation(s). In shorthand, it is often called the **strength** of the inductive argument. For the customer age example, you may state you are 50% confident in your conclusion; your degree of support is 50%. This means you believe there is a probability of 50% that the conclusion you've drawn is accurate.

degree of support (also called *inductive probability*) The degree of confidence in the conclusion resulting from the stated observation(s) for an inductive argument.

strength The degree of confidence in the conclusion.

Where does the degree of support come from? In the customer age example, we provided no basis; it was simply a subjective number. In fact, inductive arguments are often accompanied by **subjective degrees of support**—degrees of support based on opinion and lacking a statistical foundation. For example, a manager may claim "Based on the strong sales of our office, I

believe the entire company has higher sales.” When asked how sure she is of this statement, she may say that she is 98% sure. However, this again is just a subjective number, rooted in that person's perceived level of confidence.

subjective degrees of support Degrees of support based on opinion and lacking a statistical foundation.

The degree of support need not be subjective. Adding some basic assumptions to the observed information, and applying some well-known statistical formulas and theorems, will generate an objective degree of support. An **objective degree of support** has a statistical foundation, making it more credible as compared to a subjective degree of support. *All inductive arguments throughout the remainder of this book will be accompanied by an objective degree of support.*

objective degree of support A degree of support that has a statistical foundation, making it more credible as compared to a subjective degree of support.

Demonstration Problem

Classify the following arguments as examples of deductive or inductive reasoning:

1. If I leave for work at 7:30 and my commute to work is exactly one hour, I will arrive at work at 8:30.
2. Suppose the probability a customer will complete a follow-up survey is 25%. Then, the probability that all of the next four customers requested to complete the survey is 0.39%.
3. All divisions of the firm on the West Coast have shown a profit. Therefore, the firm as a whole is profitable.
4. The average rating given by a random group of respondents for the new

product design was 9 out of 10. Therefore, the general public thinks favorably of the new product design.

Answers:

1. Deductive reasoning. We are setting assumptions (leave work at 7:30; commute is one hour), and arriving at a conclusion. The implied method(s) of proof is little more than simple algebra.
2. Deductive reasoning. We are making an assumption about the probability of completing the survey, and that assumption leads to a (empirically testable) conclusion about the next four customers. Note that, if we tried to prove this statement by transposition, we would quickly realize that an implicit additional assumption is that there is no selection bias (defined later in the chapter) inherent in the “next four customers.”
3. Inductive reasoning. We are taking observed outcomes for a subset of the firm and making a general conclusion about the firm as a whole.
4. Inductive reasoning. We are taking observed ratings for a subgroup of the population and reaching a general conclusion about the entire population.

COMMUNICATING DATA 2.2

INDUCTIVE REASONING VIA CUSTOMER TESTIMONIES

Many firms sell what are known as “experience goods.” Characteristics, such as quality, for these goods can only be fully ascertained after purchase and consumption. For example, wine and skin care products are experience goods, since the flavor of the wine and the effectiveness of the product, respectively, can only be fully known after personal use. Recognizing that potential customers who have never purchased them before face uncertainty about “experience goods,” firms producing such goods often assemble customer

testimonies highlighting their merits and encourage current customers to leave positive online reviews. The hope is that potential customers will use the observed data from current customers to draw conclusions (via inductive reasoning) about product quality. An individual considering a purchase of an unfamiliar skin care product may reason, based on mostly positive testimonies about the product's effectiveness, that the product is effective and decide to make a purchase. But while firms may hope potential customers will make such an inductive argument, some potential customers may be reluctant to draw such a clean conclusion. Some may worry that those giving testimonies are a "special" or "select" group based on their prior willingness to buy the product or willingness to give a testimony. How might such "selection" lead to errant conclusions? We discuss the issue of *selection bias* in greater detail later in this chapter.

We next will demonstrate how to evaluate assumptions using inductive reasoning along with an objective degree of support. In [Chapter 3](#), we expand on these ideas, detailing in general terms how to form a conclusion about a population using a data sample, and how to calculate an objective degree of support.

EVALUATING ASSUMPTIONS

LO 2.6 Explain how inductive reasoning can be used to evaluate an assumption.

Now that we have defined inductive reasoning, we can turn to a key point of interplay between deductive and inductive reasoning. We finished our discussion of deductive reasoning by defining empirically testable conclusions, noting that we can use inductive reasoning when "testing" these conclusions and ultimately evaluate the assumptions leading to them. Now we will provide a more concrete discussion of this process of evaluating assumptions, building a basic theoretical framework on which we will expand

in the next chapter.

The basic process of evaluating assumptions is as follows. First, consider a deductive argument where an assumption(s) leads to an empirically testable conclusion, and suppose that conclusion concerns an outcome probability (or probabilities). Next, suppose we have a data sample for the outcome. We can then “test” our conclusion by comparing the observed outcomes in the data sample to their corresponding probabilities. After making this comparison, we decide whether the conclusion “passes” the test or fails; this decision is an application of inductive reasoning, as we are using the specific (the data sample) to say something about the general (our empirically testable conclusion). If the empirically testable conclusion fails the test, a simple application of transposition (Not B implies Not A) implies we must reject/reconsider at least one of the assumptions leading to this conclusion; if it passes, then the data sample does not give reason to reject/reconsider the assumption(s).

[Reasoning Box 2.2](#) summarizes the process of evaluating assumptions, and [Figure 2.4](#) illustrates it.

46

REASONING BOX 2.2

INDUCTIVE REASONING FOR EVALUATING ASSUMPTIONS

1. Suppose through deductive reasoning, we have arrived at an empirically testable conclusion.
2. Collect a data sample. Using the data sample and inductive reasoning, test the empirically testable conclusion and decide either to reject the conclusion or affirm that the data are sufficiently consistent with the conclusion.
3. If the decision is to reject the empirically testable conclusion, through transposition reject at least one of the assumptions that led to the empirically testable conclusion you rejected.

FIGURE 2.4 Evaluating Assumptions via Testing Empirically Testable Conclusions

1. Deductive reasoning generates empirically testable conclusion.



2. Inductive reasoning tests empirically testable conclusion.



3. If test fails, transposition leads to rejection/reconsideration of assumption(s).



Now that we have the basic framework for evaluating assumptions, let's consider some simple ways of applying it by revisiting our example of flipping a quarter. Supposing again we assume the quarter is fair (probability of heads equals probability of tails, 50%), we can apply some basic algebra and probability theory to arrive at an empirically testable conclusion: the probability of seeing any number of heads (0–5) for five coin flips is as described in [Table 2.1](#) (re-created below). This line of reasoning represents the first step in [Figure 2.4](#).

Next, we collect a data sample to test our conclusion. For example, the data sample may consist of five coin flips, recorded in [Table 2.2](#). Using these data, we see that the total number of heads for these five flips was 5.

TABLE 2.1 Probabilities for Varying # of Heads When Flipping a Quarter Five Times

# OF HEADS	0	1	2	3	4	5
Probability	3.125%	15.625%	31.25%	31.25%	15.625%	3.125%

TABLE 2.2 Outcomes for Five Coin Flips

FLIP	OUTCOME
1	Heads
2	Heads
3	Heads
4	Heads
5	Heads

We can use the number of heads we observe in our data sample to test our conclusion, the second step of [Figure 2.4](#). It is in this phase where we apply inductive reasoning: We must conclude something about the general (the probabilities in [Table 2.1](#)) from the specific (our data sample in [Table 2.2](#)). In our data sample we had five heads out of five flips. Speaking loosely, we may view such an outcome as “extreme” or highly unlikely, and devise the following rule: If I observe an “extreme” outcome, I will conclude the probability distribution in [Table 2.1](#) is incorrect; otherwise, I will consider the distribution sufficiently consistent with the data sample. Two outcomes stand out as candidates for the label of “extreme”—five heads and five tails, each of which has probability of 3.125%. Hence, if we label the outcome of five heads or five tails as “extreme” and apply our rule, we would reject the probabilities in [Table 2.1](#) (our empirically testable conclusion). All of this is an application of inductive reasoning. Further, we know that our rule leads to an incorrect rejection (i.e., rejecting the probabilities in [Table 2.1](#) when they are in fact correct) 6.25% of the time ($= 2 \times 3.125\%$). We calculate the objective degree of support for rejections resulting from this rule as 100% minus the rate of incorrect rejections, so in this case, $100\% - 6.25\% = 93.75\%$.

Once we have decided whether our empirically testable conclusion passes the test, we can evaluate our assumption (the quarter is fair), thus moving to the third step of [Figure 2.4](#). In our example, we decided to reject the conclusion. Therefore, assuming the methods of proof that led to the

conclusion are sound (easily verified by algebra and a statistics book), we must reject our assumption of a fair quarter. In contrast, had we observed, say, three heads in five flips, we would not have rejected our conclusion, and so would not have rejected the assumption of the quarter being fair.

In summary, for our quarter example, we executed the three steps toward evaluating our assumption (that the quarter is fair). First, we generated an empirically testable conclusion using deductive reasoning in the form of a probability table. Second, we collected a data sample and used inductive reasoning to determine whether or not to reject the probability table implied by our assumption. Third, upon rejecting the probability table, we applied transposition to reject our assumption of a fair quarter.

Many business applications of this reasoning process are even simpler than our quarter example. As an illustration, consider the Microsoft Surface, a laptop with a touch screen. Suppose some employees at Microsoft believe consumers did not place any value on the touch-screen feature of the product. That is, they would be willing to pay the exact same amount for a Surface whether or not it has a touch screen. How might we evaluate this claim? We would begin by assuming it's true and deriving an empirically testable conclusion. For example, if the claim is true, we might conclude that sales in two comparable towns will look similar, even if Town A is offered a Surface with a touch screen and Town B is offered a Surface with no touch screen. Next, we could find two comparable towns, and offer one

48

(Town A) the typical Surface and offer the other (Town B) a Surface with no touch screen at the same price. We would then collect data on sales for both towns and compare. Lastly, if sales in Town B are much lower than Town A, we reject the conclusion that sales of the two Surface versions are similar, and reject the claim that consumers don't value the touch screen.

We conclude by highlighting an apparent asymmetry in our inductive reasoning pertaining to empirically testable conclusions. Notice that, after observing our data sample, we either reject the testable conclusion or we do not reject; we never confirm the testable conclusion. Our inability to confirm the testable conclusion stems from our inability to prove the general from the

specific, an inherent feature of inductive reasoning, as noted at the beginning of this section. We can easily illustrate this limitation through our quarter example. Suppose we flipped the quarter 1,000 times and observed 500 heads. Would this be sufficient evidence to definitively state the quarter is fair? Such a data sample is certainly consistent with the quarter being fair, but it is still possible that it is not a fair coin (e.g., an “unfair” quarter with probability of flipping heads of 51% could conceivably generate this data sample).

2.3 Demonstration Problem

Suppose you have a common six-sided die, and you assume it is a fair die. That is, you assume there is equal probability of rolling any of the six numbers on it. Given this assumption, using some basic probability theory, you are able to show the probabilities in [Table 2.3](#) hold for the sum total of ten rolls. For example, ten rolls comprised of {3, 1, 3, 5, 4, 6, 2, 1, 4, 5} would have a total of 34.

TABLE 2.3 Probabilities for Sum Total of Ten Rolls of a Die

PROB. (TOTAL < 31)	PROB. (31 ≤ TOTAL ≤ 39)	PROB. (TOTAL > 39)
0.0485	0.903	0.0485

1. What is the empirically testable conclusion in the above description?
2. How would you test the empirically testable conclusion?
3. Outline the deductive and inductive reasoning you could use to evaluate the assumption that the die is fair.

Answers:

1. The empirically testable conclusion is the set of probabilities in [Table 2.3](#). These are probabilities for the sum of ten rolls in general.
2. You can take the die and roll it ten times, and use the sum total of the rolls

to test whether you believe the probabilities in [Table 2.3](#) are correct in general.

3. Roll the die ten times, and observe the total sum of the numbers rolled. Using inductive reasoning, if you observe an “extreme” outcome (which you might deem to be a total less than 31 or more than 39), you will conclude the probability distribution in [Table 2.3](#) is incorrect. Otherwise, you will consider the total sum you observed sufficiently consistent with the data sample. We know that this rule leads to an incorrect rejection (i.e., rejecting the probabilities in [Table 2.3](#) when they are in fact correct) 9.7% of the time ($= 2 \times 4.85\%$). Consequently, our objective degree of support when making this inductive argument is 90.3% ($= 100\% - 9.7\%$). Lastly, if you observe a total sum that leads you to reject [Table 2.3](#), using transposition, you reject at least one assumption leading to [Table 2.3](#). Here, the clear candidate assumption to reject is the die being fair.

SELECTION BIAS

LO 2.7 Describe selection bias in inductive reasoning.

As we have seen, inductive reasoning in the form of moving from the specific to the general is a crucial component in using data to evaluate empirically testable predictions, and ultimately our assumptions. However, there are many circumstances where the improper use of inductive reasoning can lead to inaccurate, or biased, conclusions. The sources of bias typically stem from an incomplete/inaccurate characterization of the data-generating process. Some of these sources of bias have foundations in psychology. For example, *confirmation bias* is the tendency to confirm a claim, and often occurs when survey questions are constructed in a leading way. Consider two alternative survey questions seeking the same information: “Brand X is currently the

leading brand of cookie in the market; what is your favorite brand?” vs. “Please indicate your favorite brand of cookie in the market.” Use of the first question is likely to lead to an overestimate of the popularity of Brand X, as respondents have a tendency to confirm the leading statement. As another example of a source of bias, *predictable-world bias* is the tendency to find order when none exists, and often occurs when people “read too much” into perceived patterns from random data. This type of bias is common in almost any casino, as some gamblers fallaciously “discover” predictable patterns in decks of cards and roulette wheels.

Confirmation bias and predictable-world bias are just two types of bias that can occur when using inductive reasoning. We conclude this chapter by focusing our attention on one additional source of bias, which will be most relevant for the inductive reasoning we will be using in this book. **Selection bias** is the act of drawing conclusions about a population using a selected data sample, without accounting for the means of selection. Intuitively, selection bias occurs when data are collected so as to create a “select” subgroup of the population, and observed characteristics of this subgroup may not generally apply to the entire population.

selection bias The act of drawing conclusions about a population using a selected data sample, without accounting for the means of selection.

There are myriad types of selection bias. However, it is useful to highlight two common types: (1) collector selection bias, and (2) member selection bias. *Collector selection bias* occurs when the collector selects the members of the data sample in a systematic way. For example, suppose we wanted to learn the average income of all households in Chicago, Illinois, and collected a data sample of household incomes, but only from a particular region of Chicago, say, the neighborhood of Lincoln Park. Would the average income of households in this neighborhood necessarily be informative about the average income of households in Chicago overall? As is the case with almost every city, Chicago consists of widely heterogeneous neighborhoods according to income. Our location selection, an affluent neighborhood,

Lincoln Park, could lead us to draw very inaccurate conclusions about Chicago as a whole. We might conclude from our data sample that Chicago households have much higher incomes, on average, than is actually the case.

Why might we collect data only from Lincoln Park, and not Chicago as a whole? Often, because of inconvenience or unobtainability, additional data collection is costly or impossible. Such instances can generate a particular form of collector selection bias, called *availability bias*, which occurs when the collector of the data sample selects the members of the data sample according to what is most readily available. A classic example of availability bias is the act of drawing conclusions about national sentiment on an issue (e.g., public education) by learning the opinions of one's friends. Here, the friends are most readily available, and their select status as friends may imply a skewed sentiment on this issue relative to the population.

Another familiar example of collector selection bias is the news. Critics of the media cite political ideology as a biased selection mechanism. However, the most egregious bias is sensationalism. News stories that professional media organizations disseminate are often

50

selected according to their sensational aspects. When forming opinions about society as a whole, people often rely on the news they consume, as it is readily available, and hence, availability bias compounds the emphasis on sensational events.

To see how things can go wrong in the news example, suppose we wanted to learn the relative likelihood of dying from a tornado versus diabetes using a data sample. Perhaps the most convenient sample one could collect is a sample of news stories on tornado deaths and diabetes deaths. But it is likely that a large proportion of tornado deaths are made into news stories while only a small proportion of diabetes deaths are, since the former are sensational and attract viewers. As a consequence, we may conclude from our sample of news stories that there is greater risk of dying from a tornado than from diabetes, when it is the opposite (by far). By failing to account for the collector selection bias, our inductive reasoning goes wrong.

Member selection bias occurs when potential members of the data sample

self-select into, or out of, the sample. To learn about household Internet content consumption at very high data speeds, we may seek data on Internet content consumption of households who have purchased high-speed Internet data subscriptions. The potential bias in the data sample is at least partially due to household self-selection; each household in the sample chose to purchase a high-speed data subscription. This self-selection may be indicative of different preferences among the subgroup that chose a high-speed data subscription compared to those who did not. Consequently, generalizing about the entire population using just this self-selected subgroup likely will lead to inaccurate conclusions.

As another example of member selection bias, suppose a car dealership sends out a questionnaire to each of its customers over the past year, seeking feedback on the performance of its employees. The dealership cannot force its customers to complete the survey, so inevitably there will be a response rate of less than 100%, very possibly 20% or 30%. Here there is self-selection according to willingness to complete and return the questionnaire. This willingness could be indicative of the type of car purchased, harshness/kindness when reviewing others, time of day of the car purchase, etc. The sentiment among the subgroup who responds may not be very indicative of the sentiment of the entire population of customers.

As we've shown, there are many ways selection bias can be an issue when using inductive reasoning to go from a data sample to a population. How can selection bias be avoided? An intuitive solution is to collect the data sample by randomly choosing members from the population. In fact, we will show in the next chapter that this random method of data collection is extremely useful in finding accurate relationships between sample features and features of the population.

REASONING BOX 2.3

SELECTION BIAS IN INDUCTIVE REASONING

Suppose that:

1. Through deductive reasoning, you have arrived at an empirically testable conclusion.
2. You have collected a data sample to test this conclusion.

If the members of a data sample are “select” in some way, particularly if the selection impacts the measurements used to test the conclusion, inductive reasoning based on this data sample generally will suffer from bias.

51

COMMUNICATING DATA 2.3

SELECTION BIAS IN NEWS NETWORK POLLS

The United States has several major news networks providing 24-hour news coverage, including CNN, Fox News, and MSNBC. These networks often conduct polls of their viewers to gauge sentiment on key political issues. For example, each network might conduct a poll asking whether the minimum wage should be increased. Suppose MSNBC conducts this poll, and the results for 5,000 respondents are:

MSNBC	
YES	NO
68%	32%

Based on these results, we may use inductive reasoning to generalize that, for the entire United States, citizens believe the minimum wage should be increased by nearly a 2-1 margin. However, there is a clear concern when trying to draw such a conclusion—the viewers of MSNBC are a select sample. In fact, MSNBC viewers tend to vote for the Democratic Party, which often is in favor of minimum wage increases. Consequently, using this poll to draw conclusions about the general population will almost certainly suffer from selection bias. It is entirely possible that, were the poll administered to Fox

News viewers instead, the percentages would be flipped. This is because Fox News viewers tend to vote for the Republican Party, which often is against minimum wage increases. Polls administered to viewers of any of these networks generally suffer from selection bias, and one should use severe caution in trying to use them to draw general conclusions about the population.

RISING TO THE **data**CHALLENGE

Testing for Sex Imbalance

Broadly speaking, the reasoning process you might follow to test your colleague's claim of a 3% purchase rate among women would follow [Figure 2.4](#). You would begin by deductively reasoning from the assumption of a 3% purchase rate to an empirically testable conclusion. You would then use a data sample to test the conclusion, and using inductive reasoning decide whether or not to reject the conclusion. Lastly, if you reject the conclusion, you would reject the original assumption of the 3% purchase rate among women.

For this particular problem, you could show that a purchase rate of 3% implies the following probabilities for total purchases out of 500 as shown in [Table 2.4](#). Hence, these probabilities serve as an empirically testable conclusion.

TABLE 2.4 Probabilities for Total Purchases out of 500

PROB (TOTAL < 7)	PROB (7 ≤ TOTAL ≤ 25)	PROB (TOTAL > 25)
0.007	0.988	0.005

52

You may then choose to reject these probabilities as being correct if there are fewer than 7 or more than 25 purchases among the 500 women in the data sample. For example, if you observe, say, 41 purchases in the data sample, you would reject the probabilities. You would then use transposition to reject that the rate of purchase among females in the population is 3%. It is important

to note, though, that arriving at this final conclusion relies on accepting additional, implicit assumptions, e.g., there is no selection bias in our sample. In the next chapter, we show how to make these additional assumptions explicit.

SUMMARY

This chapter introduced basic principles of reasoning in the form of deductive and inductive reasoning. It showed how we can go from general assumptions to specific conclusions (deductive reasoning), and how deductive reasoning can lead to empirically testable conclusions. We then showed how we can go from specific observations to general assertions (inductive reasoning), a process that can be used to test empirically testable conclusions, and ultimately the assumptions that lead to them. Lastly, this chapter highlighted some important sources of bias in inductive reasoning, with a focus on selection bias.

The reasoning framework developed in this chapter, and its application to data samples and populations, is a foundation for virtually every topic that follows in this book. The conclusions we will try to draw about general populations will become more formal and more intricate beginning with the next chapter; consequently, the path from data sample to conclusions about a population will be more detailed and complex. However, the underlying structure will follow very closely the framework described in this chapter.

KEY TERMS AND CONCEPTS

data sample

deductive reasoning

degree of support

direct proof

empirically testable conclusion

inductive reasoning

logic

objective degree of support

population

reasoning

robustness

selection bias

strength (of an inductive argument)

subjective degrees of support

transposition

CONCEPTUAL QUESTIONS connect

1. Provide a concise definition of reasoning. (LO1)
2. Classify the following arguments as examples of deductive or inductive reasoning: (LO5)
 - a. Most of the northern Europeans whom I have met were tall. Therefore, I believe northern Europeans tend to be tall in general.
 - b. If overeating causes weight gain, then I will weigh more after binge-eating on a cruise.
 - c. The statement “All dogs go to heaven” implies my dog will go to heaven.
 - d. Last year, firms in my business sector grew by 40% on average. Therefore, I believe the mean growth for this sector in general is 40%.

53

-
3. Consider the following statement: If demand curves are always downward-sloping and we have 4,000 sales when charging \$20 for our product, then increasing our price to \$25 will result in sales less than 4,000. (LO2)
 - a. Prove this statement using a direct proof.
 - b. Prove this statement by transposition.
 4. Identify what is wrong with the following inductive argument: “Early polling for the election is in, and we are losing by 4 points to our opponent. Based on these numbers, there is no more than a 1% chance that we win this

election.” (LO4)

5. Consider the following claim: “If we target our advertising to a younger audience, our margins will increase.” Suppose that separate analyses have convincingly shown younger customers are less price sensitive with regard to the product being sold. Explain why the ability to reason is important when assessing data-driven conclusions such as this one. (LO1)
6. Refer to the claim made in Question 5. Regardless of whether you believe the claim to be true, identify which of the following can be classified as a direct proof, transposition, or neither. (LO2)
 - a. If our margins increase, we must be selling to a less price-sensitive audience. Younger customers tend to be less price sensitive with regard to our product, so it must be that our advertising was targeting to this group.
 - b. If our margins do not increase, we must not be targeting our product to a group with lower price sensitivity. Given younger customers tend to be less price sensitive with regard to our product, we must not be targeting our advertising to a younger audience.
 - c. If we target our advertising to a younger audience, we will have more customers with lower price sensitivity, since younger customers tend to be less price sensitive with regard to our product. Consequently, we should see our margins increase.
7. Suppose a colleague is trying to convince you that the claim in Question 5 is true. (LO2)
 - a. Which method of proof is generally more effective at identifying hidden assumptions?
 - b. List one hidden assumption for the above claim.
8. Suppose your friend claims to be a faster sprinter than you are. If we assume this claim to be true, what is an empirically testable conclusion that would follow from this assumption? (LO3)
9. You are a sales manager, and you want to measure the average sales performance for new hires at your firm. To do this, you collect sales figures for each month over two years for every new hire that started in September 2015 and remained with the firm until September 2017. Explain

why using inductive reasoning to draw conclusions about the average monthly sales for new hires using these data might suffer from selection bias? (LO7)

10. Explain the difference between a subjective degree of support and an objective degree of support for an argument made via inductive reasoning. (LO4)
11. Which type of reasoning (deductive or inductive): (LO5)
 - a. Requires use of observation?
 - b. Requires assumptions?
 - c. Can vary in degree of support?
12. Which of the following is an empirically testable conclusion: (LO3)
 - a. The probability that Team A defeats Team B in a game of soccer is 65%.
 - b. The probability Bruce Lee (the late, famous martial artist) would have won the U.S. national karate championships in 1970 was 83%. (Note: Bruce Lee did not compete in any karate tournaments during his lifetime.)
 - c. The probability that store-level profits rise next year is 75%.

54

-
13. A used car salesman claims to be so good at his job that he has a 50/50 chance of selling a car to any person that walks on the lot. You note that, if this is true, the probability of making no sales to the next five customers is just 3%. Explain how you can use inductive reasoning to evaluate the car salesman's claim. (LO6)

QUANTITATIVE PROBLEMS

14. Suppose you receive an e-mail from a stock broker who claims to be able to accurately predict whether any given stock will rise or fall in price during the subsequent month. To "prove" her claim, she makes a prediction about performance (higher price or lower price) for ten stocks over the next month. You are skeptical of the broker's claim, and assume she simply guesses which stocks will improve or worsen in price over any given

month. Put another way, you assume she has a 50% chance of being correct in her prediction for any given stock. Based on this assumption, you derive the following probabilities concerning her ten picks: (LO6)

Number of correct picks	0	1	2	3	4	5	6	7	8
Probability	0.001	0.010	0.044	0.117	0.205	0.246	0.205	0.117	0.044

- a. What is the empirically testable conclusion resulting from your deductive reasoning?
 - b. How could you test your empirically testable conclusion using a data sample?
 - c. Outline the inductive and deductive reasoning you could use to evaluate whether or not the broker is simply guessing in her stock picks.
- 15.** List two ways by which the inductive reasoning you utilized in Problem 14 could suffer from selection bias. (LO7)

Reasoning from Sample to Population

3

LEARNING OBJECTIVES

After completing this chapter, you will be able to:

- LO3.1** Calculate standard summary statistics for a given data sample.
- LO3.2** Explain the reasoning inherent in a confidence interval.
- LO3.3** Construct a confidence interval.
- LO3.4** Explain the reasoning inherent in a hypothesis test.
- LO3.5** Execute a hypothesis test.
- LO3.6** Outline the roles of deductive and inductive reasoning in making active predictions.

Chapter opener image credit: ©naqiewei/Getty Image

dataCHALLENGE Knowing All Your Customers by Observing a Few

Congratulations! You just developed and launched a new app that has been heavily downloaded by smartphone users. You are offering the app for free, but plan to ultimately monetize the product by selling advertising spots. Potential advertisers are interested in the demographics of the people who download your app, to be sure they will be able to hit their desired audience. One demographic feature they care about is the mean income of your user base, and they request that you provide information about mean income to them.

In response, you ask a random subset of people downloading your app to provide their annual income. Your request results in the following dataset ([Table 3.1](#)):

56

TABLE 3.1 Users' Reported Incomes

USER ID	INCOME	USER ID	INCOME
1	\$45,711	20	\$39,684
2	\$43,840	21	\$43,765
3	\$49,810	22	\$43,333
4	\$36,382	23	\$41,591
5	\$40,192	24	\$37,194
6	\$44,712	25	\$38,004
7	\$39,497	26	\$42,380
8	\$35,921	27	\$40,881
9	\$44,189	28	\$40,344
10	\$36,614	29	\$43,025
11	\$44,520	30	\$36,976
12	\$40,269	31	\$46,013
13	\$44,689	32	\$33,448
14	\$53,520	33	\$40,061
15	\$43,830	34	\$37,455
16	\$40,593	35	\$48,375
17	\$44,831	36	\$41,791
18	\$36,692	37	\$34,148
19	\$47,117		

Using these data, what can you tell your potential advertisers about the mean income of the full group of people downloading your app? If an advertiser

claims the mean income of your user base is only \$38,000, can you refute this?

Introduction

Within the reasoning framework established in the prior chapter, this chapter describes how we can use information contained in a data sample to draw general conclusions about a population. To do so, we first introduce some basic statistical terms and concepts. Then, we describe in detail confidence intervals and hypothesis tests, highlighting how deductive and inductive reasoning play a crucial role when we try to draw general conclusions based on the data we observed. More broadly, by using a framework centered on reason, we can clearly and effectively answer questions like:

- “What must I believe in order for me to draw a general conclusion about what is going on with my company, based on what I see in a dataset?”
- “How confident am I in the conclusions I have just drawn based on the data I have seen?”
- “What is the line of reasoning that took us from this number to that conclusion?”

Distributions and Sample Statistics

LO 3.1 Calculate standard summary statistics for a given data sample.

When attempting to draw a conclusion about a population using a data sample, it is common practice to do so by drawing a conclusion about a **population parameter**, defined as a numerical expression that summarizes

some feature of the population. For example, we may want to learn about the mean income of all people in the United States. In this case, all people in the United States constitute the population, and mean income is the population parameter. As highlighted in the previous chapter, we want our conclusion(s) about this population parameter to have an objective degree of support, meaning we must utilize both inductive *and* deductive reasoning. Why deductive reasoning? Because calculating the objective degree of support relies on assumptions and methods of proof (e.g., statistical theory). The two most prominent applications of reasoning (with objective degree of support) to learn about population parameters are: (1) construction of a confidence interval and (2) hypothesis testing. However, before detailing these two methods, we present some basic information about distributions of random variables and summary statistics for data samples.

population parameter A numerical expression that summarizes some feature of the population.

DISTRIBUTIONS OF RANDOM VARIABLES

For a given population consisting of numerical values, define X_i as a **random variable** constituting a single draw (denoted “draw i ”) from that population. As a random variable, X_i can take on multiple values, with any given realization of the variable being due to chance (or randomness). In contrast, a **deterministic variable** is one whose value can be predicted with certainty. For example, define A_i as the area of any square i with side length of three centimeters. Then, we can perfectly predict that A_i will always be nine centimeters squared, meaning A_i is a deterministic variable.

random variable Variable that can take on multiple values, with any given realization of the variable being due to chance (or randomness).

deterministic variable Variable whose value can be predicted with certainty.

A random variable can be discrete or continuous. A **discrete random variable** can take on only a countable number of values (e.g., 2, 6, 15, 24, . . .), while a **continuous random variable** takes on an (uncountable) infinite number of values (e.g., all the numbers, to any decimal place, between 0 and 1). The probabilities of individual outcomes for a discrete random variable are represented by a **probability function**; for a continuous random variable, they are represented by a **probability density function (pdf)**.

discrete random variable A variable that can take on only a countable number of values.

continuous random variable A variable that takes on an (uncountable) infinite number of values.

probability function A function used to calculate probabilities of individual outcomes for a discrete random variable.

probability density function (pdf) A function used to calculate probabilities of individual outcomes for a continuous random variable.

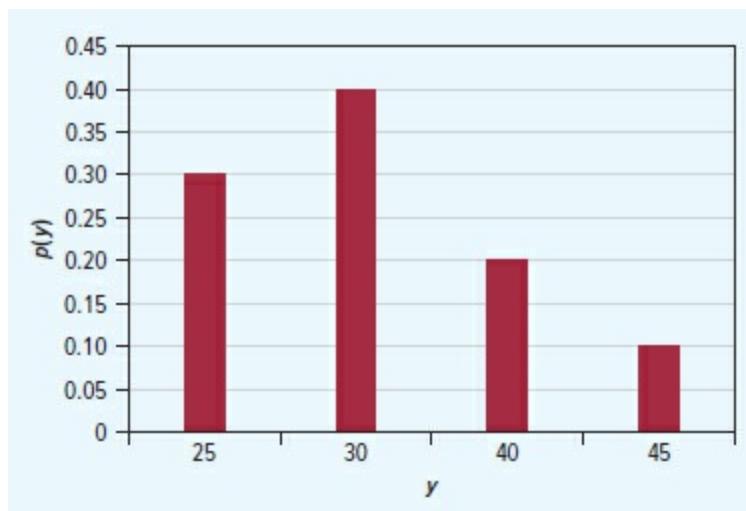
To illustrate these and subsequent definitions pertaining to random variables (and later pertaining to samples), consider the following two random variables. First, consider the ages for a group of 10 people. Suppose three of the people are 25, four are 30, two are 40, and one is 45. Then, we can define Y_i as the age for a single draw from these 10 people; thus, Y_i is a discrete random variable. Further, we write the probability function for Y_i as the probabilities for each possible outcome, i.e., the probabilities for an age of 25, 30, 40, and 45. For example, the probability that Y_i is 25 is three people out of ten ($3/10 = 0.3$), and we write this as $p(25) = 0.3$. Following this approach for all four possible outcomes, we have: $p(25) = 0.3$; $p(30) = 0.4$; $p(40) = 0.2$; $p(45) = 0.1$. [Figure 3.1](#) provides a graphical representation of this random variable.

Next, consider the height of every adult male in the United States. We

can define Z_i as a random variable consisting of the height (in inches) for a single draw from this population. Technically, Z_i is also a discrete random variable, since there are a finite number of adult males in the United States. However, since this population is so large, and height can take on many values (particularly when measured with high precision), it is common practice to treat Z_i as a continuous random variable. In fact, we can treat Z_i as a specific type of

58

FIGURE 3.1 Probability Function for Y (Age)



continuous random variable, called a **normal random variable**. A normal random variable has a “bell shaped” pdf, as illustrated in [Figure 3.2](#). The formula for the pdf shown there is:

normal random variable A specific type of continuous random variable with a bell-shaped pdf.

$$f(z) = \frac{e^{\frac{-(z-70)^2}{2 \times 2^2}}}{2\sqrt{2\pi}}$$

This formula may look a bit ugly, but it is a special case of the general pdf

formula for normal random variables, which we will present shortly.

For a normal random variable, and any other continuous random variable, the pdf allows us to calculate the probabilities that the random variable falls in various ranges. In particular, the probability that a random variable falls between two numbers A and B is the area under the pdf curve between A and B . In [Figure 3.3](#), we highlight the region representing the probability that Z_i falls between 66 and 71. Unfortunately, the process of calculating these areas under the pdf is complicated by the fact that there are no simple formulas for them; however, we can calculate all these probabilities using numerical methods, which even the most basic of today's computers can implement. We explain how to do this calculation using Excel in [Demonstration Problem 3.1](#).

FIGURE 3.2 A Normally Distributed Random Variable

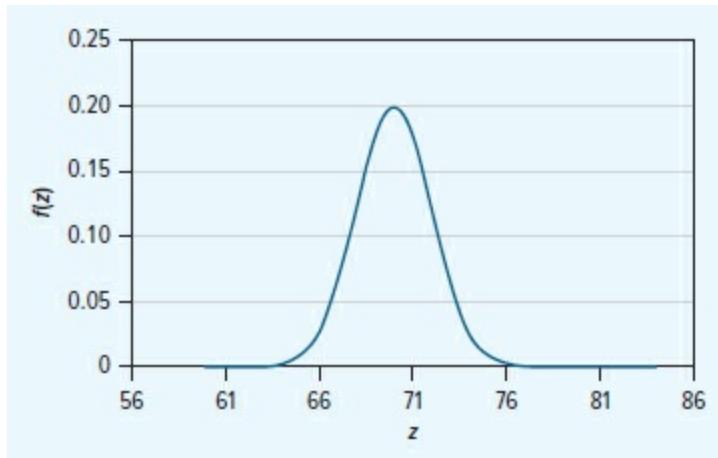
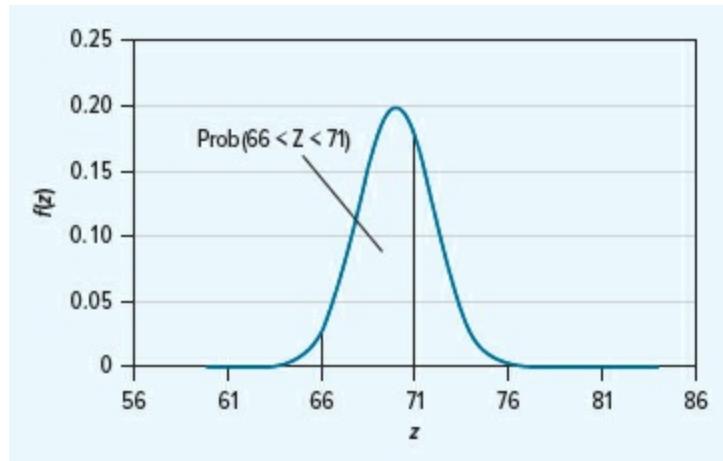


FIGURE 3.3 Probability of Z Falling between 66 and 71



3.1 Demonstration Problem

Suppose the number of daily visitors to a website is normally distributed with expected value of 2,087 and standard deviation of 316. What is the probability that the number of visitors is between 1,500 and 2,000?

Answer:

This probability is the area under the pdf for this distribution between 1,500 and 2,000. If we define X_i as a random variable representing the number of visitors for a given day for this website, then given the expected value and standard deviation of X_i , we know the pdf is:

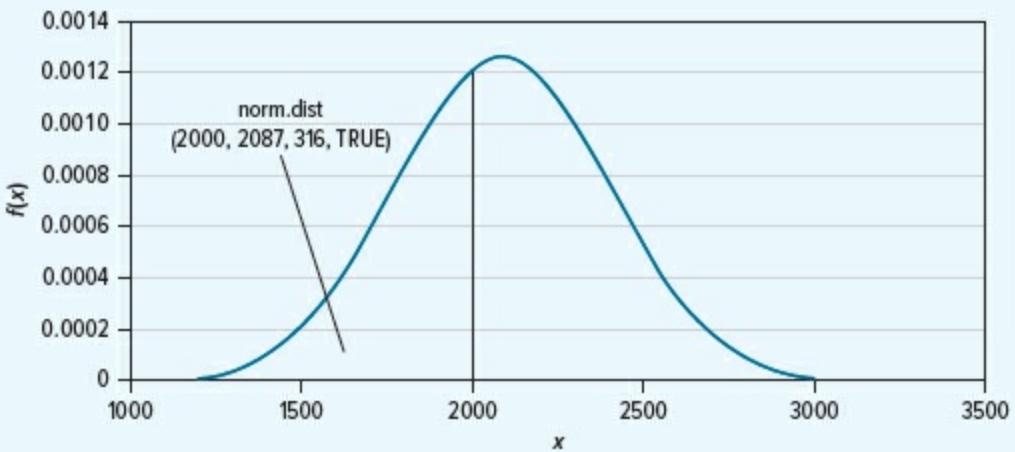
$$f(x) = \frac{e^{\frac{-(x-2087)^2}{2 \times 316^2}}}{316\sqrt{2\pi}}$$

Although we know the pdf, calculating the area under it between 1,500 and 2,000 is not a simple formula we can solve by hand; we need to call on a computer for the calculation. In Excel, we can use the formula “norm.dist.” This formula is able to calculate the area under a pdf up to a specified value for the random variable. For example, norm.dist(2000,2087,316,TRUE) will provide the probability that a normal random variable with mean 2087 and standard deviation of 316 is less than or equal to 2000. We illustrate this concept in

Figure 3.4. Note that “TRUE” indicates we want this cumulative probability; if we instead type “FALSE,” it will provide the value of the pdf when $x = 2000$. To solve our problem, we can calculate the probability that X_i is less than 2,000 and subtract the probability that X_i is less than 1,500. This will give us the probability that X_i is between 1,500 and 2,000. So, in Excel, we would use the following formula in a cell: `norm.dist(2000,2087,316,TRUE) – norm.dist(1500,2087,316,TRUE)`. If you type this in, you will find it equals approximately 0.3599, i.e., nearly 36%.

60

FIGURE 3.4 Calculation of Cumulative Probability for X



For any given distribution of a random variable, we are often interested in two key features—its center and its spread. Roughly speaking, we may ask what, on average, is the value we should expect to observe when taking a single draw for a random variable. And we may ask how varied (or spread out) we should expect the observed values to be when we take multiple draws for a random variable. A common measure for the center of a distribution is the **expected value**. We denote the **expected value** of X_i (also called the

population mean) as $E[X_i]$ and define it as the summation of each possible realization of X_i multiplied by the probability of that realization.

expected value (population mean) The summation of each possible realization of X_i multiplied by the probability of that realization.

A common measure for the spread of a distribution is the variance. We define the **variance** of X_i as $\text{Var}[X_i] = E[(X_i - E[X_i])^2]$. Another common measure of spread is the standard deviation. The **standard deviation** of X_i is simply the square root of the variance: $\text{s.d. } [X_i] = \sqrt{\text{Var } [X_i]}$. As can be seen from these formulas, the variance and standard deviation are very similar measures of spread, so why do we bother to construct both of them? The short answer is that each plays a useful role in constructing different statistics that are commonly used for data analysis. The statistics we will be using in this book generally rely on standard deviation.

variance A common measure for the spread of a distribution.

standard deviation The square root of the variance.

All three of the above features for a distribution—the expected value, variance, and standard deviation—are examples of population parameters, as each summarizes a feature of the population. Using our definitions, we can calculate each parameter for our random variables Y_i and Z_i . For Y_i , we have:

$$E[Y_i] = 0.3 \times 25 + 0.4 \times 30 + 0.2 \times 40 + 0.1 \times 45 = 32$$

$$\begin{aligned} \text{Var}[Y_i] &= 0.3 \times (25 - 32)^2 + 0.4 \times (30 - 32)^2 + 0.2 \times (40 - 32)^2 + 0.1 \times (45 \\ &\quad - 32)^2 = 46 \end{aligned}$$

$$\text{s.d.}[Y_i] = \sqrt{46} = 6.78$$

For Z_i , the calculations are analogous, but a bit more complicated. Both the

expected value and variance involve solving integrals using the pdf (i.e., calculating the area

61

under the pdf). Rather than work through these formulas, we will take a more practical approach. For discrete random variables, it is common to describe the probability function and then calculate the expected value, variance, and standard deviation using the probability function as we did previously. In contrast, for continuous random variables, it is instead common to identify the distribution (e.g., normal) and state some key population parameters (e.g., expected value, variance). Given this information, we can calculate the pdf using the population parameters. For the normal distribution, the general formula for the pdf is:

$$f(x) = \frac{e^{\frac{-(x-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}}$$

Here, μ is $E[X_i]$ and σ is s.d.[X_i]. Therefore, if we specify the expected value and standard deviation, we also know the pdf for any normal distribution. Consequently, we often see normal random variables represented as their distribution type (normal), along with their expected value and standard deviation, e.g., $X_i \sim N(\mu, \sigma)$. For Z_i , the expected value is 70 and the standard deviation is 2, and so we can write it as: $Z_i \sim N(70, 2)$. Using this information, we were able to derive the pdf— $f(z)$, which we wrote out above—and use it to calculate probabilities.

DATA SAMPLES AND SAMPLE STATISTICS

Now that we have described some fundamental characteristics of distributions, we turn to features of samples. For a random variable X_i , we define a **sample of size N** as a collection of N realizations of X_i , i.e., $\{x_1, x_2, \dots, x_N\}$. For example, consider again our random variable Z_i , the height of a man in the United States, which is normally distributed with expected value

of 70 and standard deviation of 2. Suppose now we measure the heights of five American men. In this case, we have a sample size of five (i.e., $N = 5$) for the random variable Z_i . We may represent these five measurements as a set of five numbers, e.g., $\{65, 73, 70, 62, 79\}$.

sample of size N A collection of N realizations of X_i , i.e., $\{x_1, x_2, \dots, x_N\}$.

Once we have a sample, we can calculate several **sample statistics**, defined as single measures of some feature of a data sample. As with distributions of random variables, we are often interested in the center and spread of a sample. A common measure of the center of a sample is the sample mean. The **sample mean** of a sample of size N for random variable X_i is $\bar{X} = \frac{1}{N} [x_1 + x_2 + \dots + x_N] = \frac{1}{N} \sum_{i=1}^N x_i$. The sample mean for our sample size of five for Z_i is: $\bar{Z} = \frac{1}{5} \sum_{i=1}^5 z_i = \frac{1}{5} [65 + 73 + 70 + 62 + 79] = 69.8$.

sample statistics Single measures of some feature of a data sample.

sample mean A common measure of the center of a sample.

Common measures of the spread of a sample are the sample variance and sample standard deviation. The **sample variance** of a sample of size N for random variable X_i is $S^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{X})^2$. The **sample standard deviation** of a sample of size N for random variable X_i is the square root of the sample variance: $S = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{X})^2}$. Using our sample of size five for Z_i again, we can calculate the sample variance as $S^2 = \frac{1}{5-1} \sum_{i=1}^5 (z_i - \bar{Z})^2 = \frac{1}{4} \cdot [(65 - 69.8)^2 + \dots + (79 - 69.8)^2] = 44.7$. Also, the sample standard deviation for this sample is $\sqrt{44.7} = 6.69$.

sample variance Common measure of the spread of a

sample. For a sample of size N for random variable X_i , is

$$S^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{X})^2.$$

sample standard deviation The square root of the sample variance. For a sample of size N for random variable X_i , is

$$S = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{X})^2}$$

Note that these sample statistics are clear analogs to their distributional counterparts. In particular, the way we calculate the sample mean is exactly how we would calculate

62

expected value for a discrete random variable whose possible values are the elements of the sample, each occurring with probability $\frac{1}{N}$. For our example with Z_i , the sample mean for our sample of five is exactly the same as the expected value we would calculate for a discrete random variable, call it W_i , that takes on values $\{65, 73, 70, 62, 79\}$, each with probability $\frac{1}{5}$. The calculations for the sample variance and sample standard deviation have the exact same intuitive link to variance and standard deviation, but for one adjustment. Rather than use $\frac{1}{N}$, each sample formula uses $\frac{1}{N-1}$. This adjustment ensures that the sample variance and sample standard deviation are unbiased estimates for the variance and standard deviation of the sampled random variable—a concept we describe further in the next section.

3.2

Demonstration Problem

Suppose we take a sample of 12 Google employees and collect information on their annual salaries. The data are in [Table 3.2](#). Calculate the sample mean, sample variance, and sample standard deviation for this sample.

Answer:

Let Salary_i be the salary for employee i . The sample mean is

$$\frac{1}{11} \sum_1^{12} \text{Salary}_i = \$115,020. \text{ The sample variance is}$$

$$\frac{1}{11} \sum_1^{12} (\text{Salary}_i - \$115,020)^2 = 1,145,312,577. \text{ The sample standard}$$

$$\text{deviation is } \sqrt{\frac{1}{11} \sum_1^{12} (\text{Salary}_i - 115,020)^2} = 33,482.47.$$

TABLE 3.2 Salaries of Google Employees

EMPLOYEE NUMBER	SALARY
1	\$126,971
2	\$52,895
3	\$160,888
4	\$97,278
5	\$82,866
6	\$152,594
7	\$93,694
8	\$128,556
9	\$97,875
10	\$116,972
11	\$166,538
12	\$103,109

LO 3.2 Explain the reasoning inherent in a confidence interval.

LO 3.3 Construct a confidence interval.

Confidence Interval Building on an example from [Chapter 2](#), suppose a firm wants to know the average age of all its customers, and believes it is too costly to collect this information for each and every one. Instead, it distributes to a subset of its customers a questionnaire asking their age, and ultimately collects a data sample containing the ages of 872 of its customers.

Hence, we have a sample of size $N = 872$. Using this example, we can define the following:

Age_i = a random variable defined as the age of a single customer

age_i = the observed age of customer i in the sample

$\overline{\text{Age}} = \frac{1}{872} \sum_{i=1}^{872} \text{age}_i$ = the sample mean

$s_{\text{age}} = \sqrt{\frac{1}{871} \sum_{i=1}^{872} (\text{age}_i - \overline{\text{Age}})^2}$ = the sample standard deviation

$\mu_{\text{age}} = \frac{1}{\text{Pop}} \sum_{i=1}^{\text{Pop}} \text{Age}_i$ = the population mean (the mean age of all the firm's customers, where Pop is the total number of customers)

Suppose the data are such that the sample mean is 43.61, and the sample standard deviation is 12.72.

If we use this sample to make a best guess for the mean age of all the firm's customers (the population mean), it seems obvious we should use the sample mean (43.61). In fact, the sample mean is an example of an **estimator** —a calculation using sample data that is used to provide information about a population parameter. Here, the sample mean is an estimator for the population mean. Consequently, we might inductively reason that the mean age of all the firm's customers is 43.61, based on the mean age of the sample we observed.

estimator A calculation using sample data that is used to provide information about a population parameter.

While this may seem like sound reasoning, it relies on a crucial assumption. To see this, suppose we learned that the sample we observed consisted of the firm's 872 youngest customers. In that case, would 43.61 be a good guess for the population mean? Certainly not. It must be the case that the population mean is larger than 43.61, and perhaps much larger. This hypothetical scenario highlights the fact that we are making an implicit assumption when we use the sample mean as a best guess for the population mean. In essence, we are assuming our inductive reasoning does not suffer from selection bias (introduced in [Chapter 2](#)). In practice, the standard

assumption we make is that the data sample is a random sample from the population. A **random sample** is one where every member of the population has an equal chance of being selected.

random sample A sample where every member of the population has an equal chance of being selected.

Throughout the remainder of the book, we will be working with random samples. However, it can be instructive to consider ways a data sample may not be random. For our customer-age example, suppose the firm administers the questionnaire only between 11: 00 P.M. and 1: 00 A.M. in customers' time zones. This feature of the questionnaire alone implies we do not have a random sample of customers; those who are purchasing between these two times are a select group. The consequence of this selection is intuitive; if it is mostly younger people who are awake and buying products online during that time, then the sample mean age will tend to underestimate the population mean age. There are ways to "correct" for situations where the data sample is not random, as in the previous example; however, those methods are outside the scope of this book and require additional sets of assumptions.

64

By simply assuming we have a random sample in our customer-age example, we already can see intuitively the interplay between deductive and inductive reasoning when moving from our sample to the population. In particular, we have the deductive argument that: "If we have a random sample, the sample mean is a "reasonable guess"—or more formally, an unbiased estimate (detailed below)—for the population mean." Then, we inductively reason that the population mean is 43.61 after observing this value for the sample mean. We support this inductive argument by noting that we are using an unbiased estimate.

While showing that an inductive argument is without bias does constitute support for the argument, it still falls short of providing an objective degree of support. How sure are we that 43.61 is the population mean: 80% sure . . . 50% sure . . . 10% sure? In fact, we generally will have very little certainty if we try to pin down the population mean to a single number. The likelihood

that a sample mean exactly equals its corresponding population mean is virtually zero in almost any application.

To see this, consider a simple example. Suppose the full population of a firm's customers consists of 60 people whose ages are 20, 21, 22 . . . , 77, 78, 79. Then, we know the population mean is 49.5 (taking the sum and dividing by 60). Suppose now that we take a sample of size 10 from this population; one such sample might consist of ages (22, 29, 30, 37, 48, 52, 55, 61, 65, 74), meaning the sample mean is $47.3 \neq 49.5$. If we take many more samples of size 10, their sample means will almost certainly be something other than 49.5. (Feel free to give this a try, to check the conclusion for yourself.)

Since inductive arguments about population parameters that use a single number almost certainly have little strength, we should instead consider using a range of numbers. Specifically, we can make arguments that look like: “Based on observing a sample mean of A, we conclude the population mean is between B and C.” By using a range of numbers, rather than one specific number, we need only have the population parameter fall in that range for our argument to be correct, rather than get the population parameter exactly right.

In our example using customer ages, we might make the following inductive argument: “Based on observing a sample with 872 observations, a sample mean of 43.61, and a sample standard deviation of 12.72, we conclude that the population mean is between 42.77 and 44.45.” The range of 42.77– 44.45 is called a **confidence interval**, defined as a range of values such that there is a specified probability that they contain a population parameter. In our example, by simply assuming we have a random sample, we can show that there is approximately a 95% probability that the range of 42.77– 44.45 contains the mean age of all the firm's customers. Hence, the objective degree of support for this inductive argument is 95%, and we call this a 95% confidence interval.

confidence interval A range of values such that there is a specified probability that they contain a population parameter.

How do we build confidence intervals and determine their objective degree of support? In what follows, we will show that by simply assuming

we have a random sample that is “reasonably large” (clarified below), we can build a confidence interval for the population mean using just the sample mean, sample standard deviation, and sample size. Further, we can determine the objective degree of support for the confidence interval.

For a given sample of size N , assume it is a random sample. This implies X_1, \dots, X_N , whose realizations constitute the sample, all have distributions mirroring the population distribution. Consequently, each has a mean of μ and standard deviation of σ , and all are identically distributed. The assumption of a random sample also implies that each random variable is **independent**, meaning the distribution of one random variable does not depend on the realization of another. Hence, we say that X_1, \dots, X_N are **independent and identically distributed, or i.i.d.**

independent The distribution of one random variable does not depend on the realization of another.

independent and identically distributed (i.i.d.) The distribution of one random variable does not depend on the realization of another and each has identical distribution.

65

Still assuming a random sample, it can be shown that the mean of \bar{X} (that is, the mean of the sample mean) is the population mean, μ . Mathematically, we have: $E[\bar{X}] = \mu$. Rather than provide a formal proof, consider the following intuition. Let X_i be the weight of an adult American, and suppose we weigh two Americans. Hence, we observe a value for X_1 and X_2 . Suppose we knew $E[X_i] = 180$ pounds. In words, we know the mean weight of all American adults is 180 pounds. If we then randomly pick and weigh two American adults, and average their weights, what should we expect that average to be? If, on average, the first measurement is 180 and the second measurement is 180, then we should expect the average of these two measures to be 180 (i.e., $\frac{180+180}{2} = 180$).

An estimator whose mean is equal to the population parameter it is used

to estimate is known as an **unbiased estimator**. Since the mean of \bar{X} is equal to the population mean, we say that the sample mean is an unbiased estimator for the population mean. This justifies using the sample mean as a “best guess” for the population mean and the center for any confidence interval we will use. One could also formally show that the mean of the sample variance is the variance you would calculate if you collected the entire population, i.e., the population variance formally, ($E[S^2] = \sigma^2$). In addition, the mean of the sample standard deviation is the **population standard deviation**, the square root of the **population variance** formally, ($E [S] = \sigma$). Hence, each sample measure is an unbiased estimator of its population counterpart.

unbiased estimator An estimator whose mean is equal to the population parameter it is used to estimate.

population standard deviation The square root of the population variance.

population variance The variance of a random variable over the entire population.

In order to construct a confidence interval for the population mean and know its objective degree of support, we must learn more about the distribution of the sample mean beyond just its expected value. In particular, we must know something about its standard deviation and its type of distribution. Regarding the former, the assumption that a data sample is a random sample implies the standard deviation of the sample mean is $\frac{\sigma}{\sqrt{N}}$. More succinctly, we have: s.d. $[\bar{X}] = \frac{\sigma}{\sqrt{N}}$. Intuitively, this simply means that the spread of the sample mean gets smaller as the sample size increases. One can verify this property of the standard deviation of the sample mean in practice via a simple exercise. Randomly draw two whole numbers between 0 and 100 and average them; repeat this process twenty times. Then, randomly draw ten whole numbers between 0 and 100 and average them, and repeat this process twenty times. Inevitably, you will find the second set of averages more closely centered around 50 than the first; thus, the mean of the larger

sample ($N = 10$) has less spread than the mean of the smaller sample ($N = 2$).

Regarding the distribution type of the sample mean, if we assume a random sample, and that the sample is “reasonably large” (described below), we can actually deduce that the sample mean has a normal distribution. Arriving at this conclusion depends on the use of the central limit theorem. The central limit theorem states that, for a large enough sample, often assumed to be at least $N = 30$, the mean of i.i.d. random variables is normally distributed. The central limit theorem applies to a wider range of cases than just means of a single random variable (as we will discuss later in the book), but for now, this simple application is enough to tell us that the sample mean is normal.

Thus far, we know that assuming a random sample with reasonably large $N (> 30)$ implies that the sample mean is normally distributed with mean of μ and standard deviation of $\frac{\sigma}{\sqrt{N}}$. We can write this simply as:

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{N}}\right)$$

We summarize this deductive reasoning in [Reasoning Box 3.1](#).

66

REASONING BOX 3.1

THE DISTRIBUTION OF THE SAMPLE MEAN

If a sample of size N is a random sample and N is “large” (> 30), then

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{N}}\right).$$

How does [Reasoning Box 3.1](#) help us to build confidence intervals for the population mean, μ ? By knowing the distribution of the sample mean, we can (using the probabilities for a normal distribution) calculate the probability

that the sample mean falls within any given distance from its mean, μ . In this context, distance is generally measured in terms of standard deviations. For example, we might measure the probability that a random variable falls within 2.7 standard deviations of its mean.

As mentioned previously, calculating probabilities for normal random variables can be complicated (e.g., requiring computer calculations). However, there are a few probabilities worth committing to memory, given their high level of utilization. Specifically, we know that for any normal random variable, it will:

- Fall within 1.65 standard deviations of its mean approximately 90% of the time.
- Fall within 1.96 standard deviations of its mean approximately 95% of the time.
- Fall within 2.58 standard deviations of its mean approximately 99% of the time.

The above cutoffs are worth committing to memory. However, it is useful to note that they are reasonably close to cutoffs of 1.5, 2, and 2.5, respectively. Although these rougher cutoff values are less precise and should not be used in general, they can be helpful when quickly and roughly assessing statistical outputs if the more precise cutoffs do not immediately come to mind.

Since the sample mean is normally distributed when there is a large, random sample, the above probabilities apply. We can write these probabilities more formally and succinctly for the sample mean as follows.

For a random data sample of size $N > 30$,

$$\Pr\left(\bar{X} \in \left[\mu \pm 1.65 \left(\frac{\sigma}{\sqrt{N}}\right)\right]\right) \approx 0.9$$

$$\Pr\left(\bar{X} \in \left[\mu \pm 1.96 \left(\frac{\sigma}{\sqrt{N}}\right)\right]\right) \approx 0.95$$

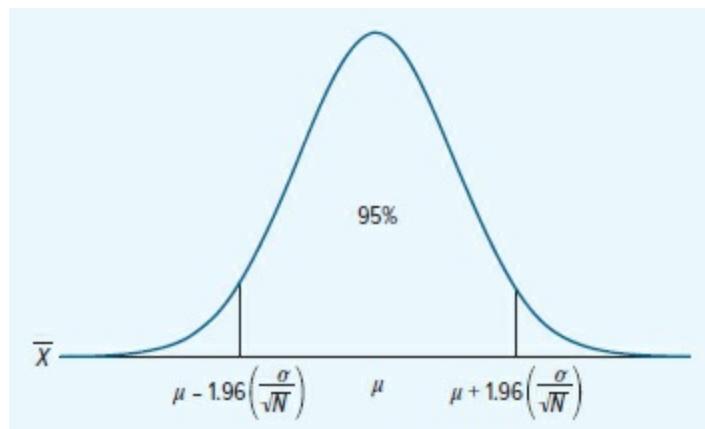
$$\Pr\left(\bar{X} \in \left[\mu \pm 2.58 \left(\frac{\sigma}{\sqrt{N}}\right)\right]\right) \approx 0.99$$

This idea is illustrated in [Figure 3.5](#) for the case of 95% probability. (For 90%, the range is narrower, and for 99% the range is wider.)

For the above probabilities, we are taking the population mean and creating an interval around it to capture a certain percentage (e.g., 90%, 95%) of all possible draws for the sample mean. For example, we have that, given a population mean of, say, 10, the sample mean will fall within 1.96 standard deviations of 10 about 95% of the time.

Although we now have substantial information about the sample mean, our objective is not to predict where the sample mean will fall; we observe the sample mean when we collect a data sample. Instead, we want to take an observed sample mean and create an interval

FIGURE 3.5 Probability Sample Mean within 1.96 Standard Deviations of Population Mean



around it that will capture the population mean with some known probability. For example, we want to find a number K such that, given a sample mean of, say, 15, the interval of $(15 \pm K \text{ standard deviations})$ will capture the population mean about 95% of the time. Fortunately, using some simple algebra, we can show the given rule implies the following:

- The sample mean plus or minus 1.65 standard deviations will contain the population mean approximately 90% of the time.

- The sample mean plus or minus 1.96 standard deviations will contain the population mean approximately 95% of the time.
- The sample mean plus or minus 2.58 standard deviations will contain the population mean approximately 99% of the time.

We summarize this rule as follows. For a random sample of size $N > 30$,

$$\Pr \left(\mu \in \left[\bar{X} \pm 1.65 \left(\frac{\sigma}{\sqrt{N}} \right) \right] \right) \approx 0.9$$

$$\Pr \left(\mu \in \left[\bar{X} \pm 1.96 \left(\frac{\sigma}{\sqrt{N}} \right) \right] \right) \approx 0.95$$

$$\Pr \left(\mu \in \left[\bar{X} \pm 2.58 \left(\frac{\sigma}{\sqrt{N}} \right) \right] \right) \approx 0.99$$

At this point, we now have formulas for confidence intervals that have objective degrees of support. For example, if we take the sample mean and then add and subtract 1.96 standard deviations, we know this will contain the population mean approximately 95% of the time. Or, put another way, we are 95% confident this interval will contain the population mean.

There remains but one problem. These formulas use the population standard deviation (σ), and we do not observe this in our data. The solution, as might be guessed, is to replace the population standard deviation with the sample standard deviation (S) in the formulas. As noted, the sample standard deviation is an unbiased estimator for the population standard deviation, justifying it as a “best guess” for σ .

Does replacing σ with S have any consequence for our confidence interval formulas? Our use of 1.65, 1.96, and 2.58 standard deviations when constructing these confidence intervals was based on probabilities for the normal distribution. When we replace σ with S , these numbers must instead be based on what's known as a *t-distribution*. The *t*-distribution generally

68

looks like a normal distribution but requires larger numbers of standard deviations to achieve the same probabilities as the normal distribution (e.g., 2, 2.5, and 4 standard deviations might be necessary to attain 90%, 95%, and

99% probabilities, respectively). However, when $N > 30$ (which we already are assuming in order to apply the central limit theorem), the difference between a t -distribution and normal distribution becomes trivial, and so the same probability formulas apply. Consequently, the confidence interval formulas become as follows. For a random sample of size $N > 30$,

$$\Pr \left(\mu \in \left[\bar{X} \pm 1.65 \left(\frac{s}{\sqrt{N}} \right) \right] \right) \approx 0.9$$

$$\Pr \left(\mu \in \left[\bar{X} \pm 1.96 \left(\frac{s}{\sqrt{N}} \right) \right] \right) \approx 0.95$$

$$\Pr \left(\mu \in \left[\bar{X} \pm 2.58 \left(\frac{s}{\sqrt{N}} \right) \right] \right) \approx 0.99$$

Revisiting our initial example involving mean customer ages, we can construct all three confidence intervals immediately using the provided sample information. Therefore, using the sample mean of 43.61, sample standard deviation of 12.72, and sample size of 872, we have the following confidence intervals for the mean age of all the firm's customers:

$$90\% \text{ confidence interval: } \left(43.61 \pm 1.65 \left(\frac{12.72}{\sqrt{872}} \right) \right) = (42.90, 44.32)$$

$$95\% \text{ confidence interval: } \left(43.61 \pm 1.96 \left(\frac{12.72}{\sqrt{872}} \right) \right) = (42.77, 44.45)$$

$$99\% \text{ confidence interval: } \left(43.61 \pm 2.58 \left(\frac{12.72}{\sqrt{872}} \right) \right) = (42.50, 44.72)$$

We summarize the reasoning associated with confidence intervals in [Reasoning Box 3.2](#).

REASONING BOX 3.2

CONFIDENCE INTERVALS

Deductive reasoning:

IF:

1. A data sample is random.

2. The size of the data sample is greater than 30.

THEN:

The interval consisting of the sample mean plus or minus 1.65 (1.96, 2.58) standard deviations of the sample mean will contain the population mean approximately 90% (95%, 99%) of the time.

Inductive reasoning: Based on the observation of the sample mean, \bar{X} , the sample standard deviation, S , and the sample size, N , the population mean is contained in the interval $\left(\bar{X} \pm 1.65 \left(\frac{S}{\sqrt{N}}\right)\right)$. The objective degree of support for this inductive argument is 90%. If we instead use the intervals $\left(\bar{X} \pm 1.96 \left(\frac{S}{\sqrt{N}}\right)\right)$ and $\left(\bar{X} \pm 2.58 \left(\frac{S}{\sqrt{N}}\right)\right)$, the objective degree of support becomes 95% and 99%, respectively.

3.3

Demonstration Problem

Your employer, who sells used cars, asks you to assess the average age of cars currently being driven in your neighborhood. To do this, you observe cars in the parking lot of a large company in your neighborhood, noting the age of each one. The data you collect are shown in [Table 3.3](#).

Build a confidence interval for the mean age of all cars in your neighborhood using these data. Supply the necessary reasoning that leads to your confidence interval. Do you believe the assumption(s) used to build your confidence interval hold in this case?

Answer:

You are 90% confident that the mean age of all cars in your neighborhood is between 7.22 and 9.37 years.

You are 95% confident that the mean age of all cars in your neighborhood is between 7.01 and 9.57 years.

You are 99% confident that the mean age of all cars in your neighborhood is between 6.61 and 9.98 years.

The reasoning is as follows: Assuming the data are a random sample, and given the sample size is more than 30, you conclude, with 90/95/99% confidence that the mean age of all cars in the neighborhood is within 1.65/1.96/2.58 standard deviations of the sample mean.

A concern with this argument may be whether the data sample is a random sample. You should ask whether cars in the lot you observed may tend to be older or younger than the cars in the neighborhood, on average.

TABLE 3.3 Age of Cars in Parking Lot

CAR NUMBER	AGE	CAR NUMBER	AGE	CAR NUMBER	AGE
1	15	21	6	41	13
2	5	22	14	42	11
3	7	23	8	43	0
4	3	24	15	44	8
5	4	25	1	45	1
6	8	26	4	46	16
7	16	27	4	47	10
8	11	28	7	48	7
9	14	29	7	49	13
10	13	30	15	50	16
11	8	31	0	51	0
12	11	32	15	52	4
13	11	33	8	53	13
14	3	34	10	54	2
15	14	35	12	55	11
16	8	36	13	56	7
17	5	37	7	57	0
18	14	38	1	58	4
19	10	39	12		
20	0	40	6		

COMMUNICATING DATA 3.1

WHAT CAN POLITICAL POLLS TELL US ABOUT THE GENERAL POPULATION?

During any election year, voters are inundated with political polls attempting to determine which candidate is currently in the lead, and by how much. For example, a polling agency may claim that a given candidate—say, Kate Sciarras—is supported by 62% of the population. To get this figure, the agency did not ask every person in the voting population whom they support. Rather, they collected a data sample and calculated the sample mean and sample standard deviation.

While often ignored when reported in the media, the validity of the sample mean as an unbiased guess for the population mean relies on us making the implicit assumption that the sample is random. Further, we know that 62% almost certainly is not an exactly correct guess for the population mean. Consequently, we use the sample to build a confidence interval. In practice, you will often see the sample mean reported with a “margin of error,” meaning the amount this figure may differ from the population mean in either direction. Since the standard confidence level is 95%, this margin of error is an approximate calculation of 1.96 times the standard deviation of the sample mean $\left(1.96 \left(\frac{s}{\sqrt{N}}\right)\right)$. For our example, the margin of error might be 3%, meaning we are 95% confident that the proportion of voters supporting Kate Sciarras is between 59% and 65%. This information is often conveyed in short-hand as: “Support is 62% with a 3% margin of error.”

LO 3.4 Explain the reasoning inherent in a hypothesis test.

LO 3.5 Execute a hypothesis test.

Hypothesis Testing Consider again our example where a firm wants to know the average age of all its customers. By building a confidence interval, we have a range of plausible values for this population parameter. However, this is not the only way to provide information about a population parameter using a data sample. On many occasions, we have in mind a value for the population parameter, and we wish to see if this value is plausible given the data we observe.

For example, a manager at the firm may believe the average age of the firm's customers is 44.5. How do we decide whether or not to believe this claim? To make our decision, we can use the ages of the customers whom we observe in our data sample to determine whether this claim is reasonable. Here, the proposal that the average age of all customers is 44.5 constitutes a hypothesis, and we wish to test this hypothesis using a data sample.

A **hypothesis test** is the process of using sample data to assess the credibility of a hypothesis about a population. Suppose again that we have a data sample with 872 customers' ages, where the sample mean is 43.61 and the sample standard deviation is 12.72. Based on this information, is the hypothesis that the mean age in the population is 44.5 credible, or should we reject this idea? We know that if the data sample is a random sample, then the sample mean is an unbiased estimator of the population mean. Consequently, 43.61 is our best guess for the population mean. However, we know the sample mean almost never exactly equals the actual population mean, so no matter how close the sample mean is to the hypothesized population mean, a small difference between the two will not be sufficient for us to conclude the hypothesized value is correct.

hypothesis test The process of using sample data to assess the credibility of a hypothesis about a population.

In contrast, we can ask whether 43.61, our best guess for the population mean, is “too far” from 44.5, and hence too unlikely to occur, for us to believe that 44.5 is a credible value for the population mean. Conducting a hypothesis test consists of making such an assessment—that is, choosing between (1) rejecting a hypothesis as

noncredible, or (2) failing to reject the hypothesis. In our example, we must choose between rejecting 44.5 as a noncredible value for the population mean, or failing to reject 44.5 as noncredible. If we reject, then we are concluding that 44.5 is not the population mean; if we fail to reject, we believe it is plausible for 44.5 to be the population mean but isn't necessarily correct.

To see how a hypothesis test works, consider again the case where we have a sample of size N , consisting of N realizations of the random variable X_i , i.e., $\{x_1, x_2, \dots, x_N\}$. The sample mean is \bar{X} , and the sample standard deviation is S . Further, the population mean is μ and the population standard deviation is σ . From [Reasoning Box 3.1](#), we know:

If a sample of size N is a random sample and N is “large” (> 30), then

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{N}}\right).$$

For a hypothesis test, we hypothesize a value for a population parameter; in our setting, we are hypothesizing a value for the population mean. This hypothesis about the population parameter's value is called the **null hypothesis**, defined as the hypothesis to be tested using a data sample. We write this as follows, where K is the hypothesized value for the population mean:

null hypothesis The hypothesis to be tested using a data sample.

$$H_0 : \mu = K$$

Our objective then is to determine whether we believe this null hypothesis is credible given the data we observe.

When building a confidence interval, we take what we observe in the

sample (e.g., sample mean, sample size, sample standard deviation) and determine what we believe is feasible for a population parameter (e.g., population mean). In contrast, when conducting a hypothesis test, we make an assumption about the population (null hypothesis), and use what we observe in the sample to assess whether this assumption is credible. Therefore, we add the null hypothesis to our set of assumptions. Doing so yields a simple expansion of [Reasoning Box 3.1](#) as shown in [Reasoning Box 3.3](#).

By assuming the null hypothesis, we know the center of the distribution for the sample mean—it is centered at K . Because the sample mean is normally distributed, we know that it will:

- Fall within 1.65 standard deviations of K approximately 90% of the time.
- Fall within 1.96 standard deviations of K approximately 95% of the time.
- Fall within 2.58 standard deviations of K approximately 99% of the time.

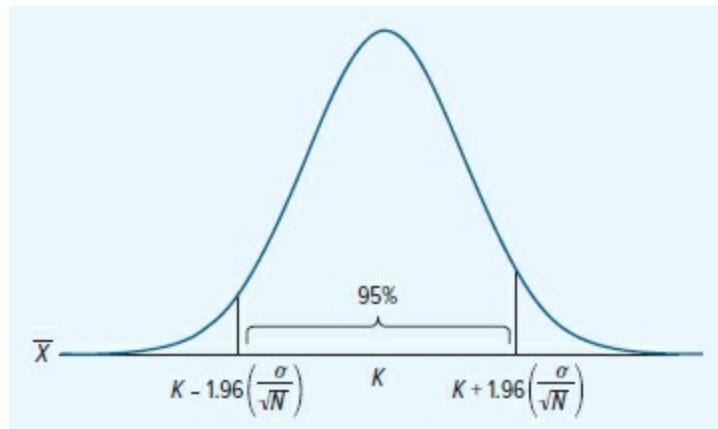
We can again write this rule more formally and succinctly for the sample mean with hypothesized population mean of K as follows.

REASONING BOX 3.3

THE DISTRIBUTION OF THE SAMPLE MEAN FOR HYPOTHESIZED POPULATION MEAN

If a sample of size N is a random sample, N is “large” (> 30), and $\mu = K$, then $\bar{X} \sim N\left(K, \frac{\sigma}{\sqrt{N}}\right)$.

FIGURE 3.6 Probability Sample Mean within 1.96 Standard Deviations of Hypothesized Population Mean (K)



For a random data sample of size $N > 30$, and population mean = K ,

$$\Pr \left(\bar{X} \in \left[K \pm 1.65 \left(\frac{\sigma}{\sqrt{N}} \right) \right] \right) \approx 0.9$$

$$\Pr \left(\bar{X} \in \left[K \pm 1.96 \left(\frac{\sigma}{\sqrt{N}} \right) \right] \right) \approx 0.95$$

$$\Pr \left(\bar{X} \in \left[K \pm 2.58 \left(\frac{\sigma}{\sqrt{N}} \right) \right] \right) \approx 0.99$$

This idea is illustrated in Figure 3.6 for the case of 95%. In contrast to Figure 3.5, the sample mean centers on an assumed value of K , rather than an unknown value of μ .

Notice that, by assuming our null hypothesis along with a large random sample, we have arrived at an empirically testable conclusion, i.e., a random variable with known distribution and consequently known probabilities for various outcomes. In fact, hypothesis tests fall exactly into the reasoning framework described in Chapter 2 for evaluating assumptions. Here, the null hypothesis is the assumption we would like to evaluate. Recall that the general process for evaluating assumptions is to: (1) Use deductive reasoning to arrive at an empirically testable conclusion, (2) Collect a data sample and use inductive reasoning to decide whether or not to reject the empirically

testable conclusion, and (3) If you reject, use transposition to reject at least one assumption. To this point, we have accomplished the first step, as we have an empirically testable conclusion in the form of the distribution of the sample mean. The remainder of this section details how to execute the remaining two steps.

The next step in conducting our hypothesis test is to collect a data sample and calculate the sample mean. Upon observing the sample mean, we then must decide whether or not to reject our deduced distribution for the sample mean from [Reasoning Box 3.3](#) $(\bar{X} \sim N(K, \frac{\sigma}{\sqrt{N}}))$. This decision hinges on whether we deem the observed sample mean as “reasonably likely,” and the probabilities we calculated above give us exactly what we

73

need to make that decision. Specifically, we can use those probabilities to make the following inductive arguments:

1. Reject the distribution if the sample mean is more than 1.65 standard deviations from K ; fail to reject otherwise. This generates a degree of support of approximately 90%.
2. Reject the distribution if the sample mean is more than 1.96 standard deviations from K ; fail to reject otherwise. This generates a degree of support of approximately 95%.
3. Reject the distribution if the sample mean is more than 2.58 standard deviations from K ; fail to reject otherwise. This generates a degree of support of approximately 99%.

Note that all three arguments are valid; changes in the criterion for rejecting the distribution are accompanied by changes in the degree of support. So, if you want very strong support when you reject, you must have a very strict criterion for rejecting, and vice versa. A common degree of support expected for such tests is 95%, and so the criterion of 1.96 standard deviations is often used.

Executing the inductive argument requires two basic steps. First, you must choose the desired degree of support (e.g., 90%, 95%, or 99%) for your

inductive argument. Once you have made this choice, the criterion for rejection immediately follows using the above figures (1.65, 1.96, or 2.58 standard deviations). The second step in executing the argument simply involves measuring how many standard deviations the sample mean is from the hypothesized population mean. If we call this difference z , we can write it as follows:

$$z = \frac{\bar{X} - K}{\sigma / \sqrt{N}}$$

Note that, to calculate z , we take the difference between the sample mean and the hypothesized population mean ($\bar{X} - K$). Then, we take that difference and divide it by the standard deviation of the sample mean (σ / \sqrt{N}). For example, suppose $\bar{X} = 11$, $K = 6$, $\sigma = 20$, and $N = 100$. Then, the sample mean differs from the hypothesized population mean by 5 ($11 - 6$). And this difference is 2.5 standard deviations $\left(\frac{5}{20 / \sqrt{100}} = \frac{5}{2} = 2.5 \right)$. If we choose degree of support 90%, we have $|2.5| > 1.65$, so we reject. In contrast, if we choose degree of support 99%, we have $|2.5| < 2.58$, meaning we fail to reject.

Unfortunately, we cannot calculate z using our sample. We observe \bar{X} and N , and have assumed a value for K , but we do not know σ . However, just as with the confidence interval, we can replace the population standard deviation (σ) with the sample standard deviation (S). By making this substitution, we can rename the difference to be t , also called the ***t-stat***, with the following formula:

t-stat The difference between the sample mean and the hypothesized population mean ($\bar{X} - K$) divided by the sample standard deviation (S / \sqrt{N}).

$$t = \frac{\bar{X} - K}{S / \sqrt{N}}$$

Here, the t -stat is an example of a **test statistic**, defined as any single value derived from a sample that can be used to perform a hypothesis test.

test statistic Any single value derived from a sample that can be used to perform a hypothesis test.

Does replacing σ with S affect our cutoff rules? That is, does using a cutoff of 1.65, 1.96, and 2.58 still correspond to a degree of support of 90%, 95%, and 99%, respectively? While these cutoffs were determined using the normal distribution, the appropriate cutoffs using the

74

t -stat come from the t -distribution. However, for a given degree of support, the cutoffs for the t -distribution closely resemble those for the normal distribution when $N > 30$. Consequently, we can simply calculate the t -stat using our sample, and compare to the same cutoffs, depending on the degree of support we've chosen. For our customer age example, the t -stat is:

$$t = \frac{43.61 - 44.5}{12.72 / \sqrt{872}} = -2.07.$$

Here, $|t| > 1.96$, so we reject our empirically testable conclusion, i.e., we reject that $\bar{X} \sim N\left(44.5, \frac{12.72}{\sqrt{872}}\right)$ with 95% confidence. Had we required 99% confidence in our inductive argument, then we would fail to reject, since $|t| < 2.58$.

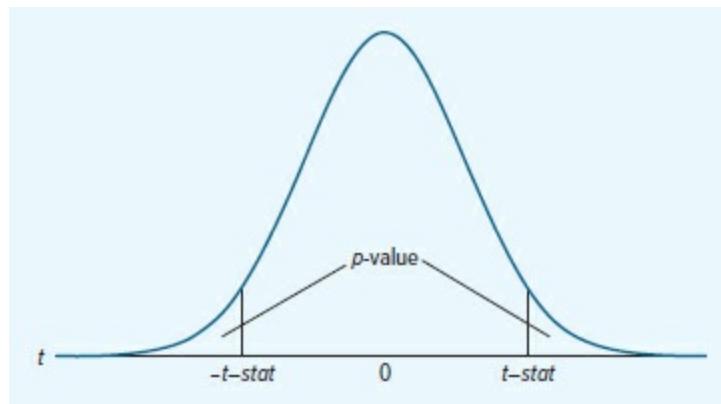
Upon observing an “unlikely” t -stat, it can be tempting to jump right to rejecting the null hypothesis. In fact, it is common to see this done in practice. However, to do so is to skip the last step of the reasoning process when evaluating an assumption: It is to skip transposition. Our unlikely t -stat only leads us to reject our deduced distribution for the sample mean, and recall that we arrived at this distribution by assuming our null hypothesis *and* a large random sample. Hence, our rejection of this distribution means only that we must reject at least one of these assumptions. Of course, it is easy to verify whether the sample is “large,” so we are left with our null and the sample being random as the assumptions we might reject. Consequently, it is

only when we are confident that we have a random sample that we can ultimately reject the null (through transposition) when we observe an unlikely t -stat. In such cases, it is not necessary to explicitly spell out this last step in rejecting the null. However, if we are not confident the sample is random, it could be that the null is in fact true but there was some form of selection when collecting the sample.

We conclude this section by presenting a common alternative method for executing the inductive argument for a hypothesis test. For the above approach, we first chose a degree of support from among 90%, 95%, and 99%, then determined the corresponding cutoff, and finally calculated the t -stat, which would ultimately be compared to a cutoff of 1.65, 1.96, or 2.58, respectively. As an alternative, we again choose a degree of support from among 90%, 95%, and 99% and calculate the t -stat. However, rather than compare the t -stat to one of our cutoffs, we instead calculate its ***p-value***, defined as the probability of attaining a test statistic at least as extreme as the one that was observed. In our case, the p -value is the probability of seeing a t -stat at least as large (in absolute value) as the one we actually observed. This concept is illustrated in [Figure 3.7](#) for the case of a positive t -stat.

p-value The probability of attaining a test statistic at least as extreme as the one that was observed.

FIGURE 3.7 Graphical Illustration of a *P*-value



The t -stat is just an observed value from a t -distribution, a well-known distribution that resembles a normal distribution and is centered at zero. In fact, for $N > 30$, the difference between a t -distribution and standard normal distribution becomes small. Consequently, using probability tables for the standard normal distribution, or more commonly algorithms executed by a computer, we can calculate the p -value for any t -stat. For example, in Excel we can calculate a p -value using the formula: $2 \times (1 - \text{norm.s.dist}(|t\text{-stat}|, \text{true}))$. This formula calculates the probability of being in the right tail (see [Figure 3.7](#)) and simply doubles it, since the distribution is symmetric. Using this approach for our customer age example means we use the formula $2 \times (1 - \text{norm.s.dist}(2.06, \text{true}))$, which equals 0.039. This means that the probability of observing a t -stat greater than 2.06 (in absolute value) is 0.039.

How unlikely our observed t -stat is if the deduced distribution for the sample mean is accurate. So, if the observed t -stat is very unlikely (i.e., has a low p -value), then we should be inclined to reject this distribution, and vice versa. This process again involves a comparison to a cutoff. Here, the cutoff indicates what we deem “unlikely enough” for us to reject: If the p -value is less than the cutoff, we reject, and fail to reject otherwise.

The cutoffs using p -values directly correspond to the degree of support you have chosen for your inductive argument. If your chosen degree of support is $D\%$, then the cutoff is $100 - D\%$, or in decimal form, $1 - D/100$. The reasoning behind this relationship is as follows. Suppose your degree of support is 95%. Then, the rule is to reject when the p -value is less than 5%, or 0.05 ($1 - 95/100$), and fail to reject otherwise. Notice that rejections will be incorrect 5% of the time using this rule. This is because 5% of the time you will observe a p -value less than 0.05 even though the deduced distribution for the sample mean is correct. Consequently, you should have 95% confidence in your inductive argument when following this rule.

Since the standard degrees of confidence used are 90%, 95%, and 99%, the standard cutoffs using p -values are 0.10, 0.05, and 0.01. Then, we can summarize the inductive arguments as follows:

1. Reject the distribution if the p -value is less than 0.10; fail to reject

otherwise. This generates a degree of support of 90%.

2. Reject the distribution if the p -value is less than 0.05; fail to reject otherwise. This generates a degree of support of 95%.
3. Reject the distribution if the p -value is less than 0.01; fail to reject otherwise. This generates a degree of support of 99%.

A convenient feature of p -values is that they allow us to choose alternative degrees of support, and easily calculate new cutoffs. For example, if we want a degree of support of 92%, then the cutoff for the p -value is $1 - 92/100 = 0.08$. In fact, the p -value tells us explicitly the degree of support we have if we choose to reject. If the p -value is 0.12, we know a rejection has degree of support 88%; if the p -value is 0.16, we know a rejection has degree of support 84%. Primarily for this reason, the use of p -value cutoffs rather than t -stat cutoffs is typically preferred when making inductive arguments in practice.

76

REASONING BOX 3.4

HYPOTHESIS TESTING

Deductive reasoning:

IF:

1. A data sample is random.
2. The size of the data sample is greater than 30.
3. The population mean is K .

THEN:

The sample mean is distributed as $\bar{X} \sim N\left(K, \frac{\sigma}{\sqrt{N}}\right)$, and will fall within 1.65 (1.96, 2.58) standard deviations of K approximately 90% (95%, 99%) of the time. This also means that the sample mean will differ by more than 1.65 (1.96, 2.58) standard deviations from K (in absolute value) approximately 10% (5%, 1%) of the time. This constitutes an empirically testable

conclusion.

Inductive reasoning:

Using t-stats. If the absolute value of the t -stat $\left(= \left| \frac{\bar{X} - K}{S/\sqrt{N}} \right| \right)$ is greater than 1.65 (1.96, 2.58), reject the deduced (above) distribution for the sample mean. Otherwise, fail to reject. The objective degree of support for this inductive argument is 90% (95%, 99%).

Using p-values. If the p -value for this t -stat is 0.0000005, reject the deduced (above) distribution for the sample mean. Otherwise, fail to reject. The objective degree of support for this inductive argument is 90% (95%, 99%).

Transposition:

If inductive reasoning leads to a rejection of the distribution for the sample mean, reject at least one of the assumptions (1, 2, or 3 above) leading to that distribution. If the sample is large, and there is confidence in a random sample, this means rejection of the null hypothesis.

3.4

Demonstration Problem

You have purchased a dataset that tracks Internet usage of 2,000 randomly chosen U.S. adults for an entire week. Among the information collected is the amount of time spent on the Internet for the week by each person. Your boss believes Internet usage averages two hours per day, or 14 hours per week. Using the following sample statistics from your data, what can you say about this claim?

Sample mean hours per week: 13.8

Sample standard deviation of hours per week: 4.8

Answer:

Using these data, the associated t -stat is -1.86 . This exceeds 1.65 (in absolute value), and so (assuming a random sample) you reject 14 as the mean hours per week for all U.S. adults with 90% confidence. Note that if you required a 95% confidence level to reject, you would fail to reject 14 as the mean weekly usage of all U.S. adults. Alternatively, the p -value for the t -stat of -1.86 is 0.063 . This is less than 0.10 , so you reject 14 as the mean hours per week for all U.S. adults with 90% confidence. Again, note that $0.063 > 0.05$, so you fail to reject 14 as the mean weekly usage of all U.S. adults if you required a 95% confidence level.

COMMUNICATING DATA 3.2

DOES WORKING AT WORK MAKE A DIFFERENCE?

In 2013, the CEO of Yahoo!, Marissa Mayer, issued an order for the company banning the practice of working from home. This was a controversial move, instigating a great deal of theoretical debate about its consequences for various outcomes, including productivity. Measuring productivity at Yahoo! is a complicated matter, but suppose a smaller company, Cincy Sales, is focused on sales and is considering following Mayer's plan. Cincy Sales contacts a consulting firm, EVConsulting, to gain insights on the possible consequences of this policy on sales. EVConsulting has collected sales data from a random subset of employees at sales firms that implemented the "no working at home" rule starting in 2017. EVConsulting creates a random variable defined as:

$$Diff_i = \text{Sales } 2017_i - \text{Sales } 2016_i,$$

In words, $Diff_i$ is the change in sales for employee i between 2016 and 2017. EVConsulting has a sample of 2,000 employees, and for that sample, it calculates the sample mean (\bar{Diff}) and sample standard deviation (S). What

is the hypothesis test that EVConsulting could run using these data that might be useful to Cincy Sales?

With these data, EVConsulting could learn about the population mean of $Diff_i$ —call this μ . That is, for all sales employees experiencing this policy change, what was their average change in sales? The hypothesis test would look as follows. The null hypothesis is that there is no change:

$$H_0: \mu = 0$$

The t -stat using the sample is:

$$t = \left| \frac{\overline{Diff} - 0}{S/\sqrt{2000}} \right|$$

Hence, if $\overline{Diff} = 500$ and $S = 22,000$, then $t = 1.02$. The p -value is 0.31. Therefore, with these sample statistics, we would fail to reject that the sales difference in the population between 2016 and 2017 for sales firms adopting the new policy is zero. We would reach this conclusion using any of the standard degrees of support (90%, 95%, and 99%).

Note that this test tells us something about the change in sales for the population of sales firms that adopted this policy. This is an interesting insight, but does knowing the mean change in sales for this population tell us the causal effect of this policy? The next several chapters will help us to think carefully about such a question.

The Interplay Between Deductive and Inductive Reasoning in Active Predictions

LO 3.6 Outline the roles of deductive and inductive reasoning in making active predictions.

We conclude this chapter by considering an overview of the roles of deductive and inductive reasoning in forming active predictions. As we

discussed in [Chapter 1](#), active predictions rely on a (measured) causal relationship between two variables (i.e., a change in X causes a change in Y). For example, we may believe the causal relationship between two variables, X and Y , is such that a one unit increase in X causes a 0.5 unit increase in Y . Assuming this relationship holds, we can then make various predictions concerning possible strategic changes in X . In particular, we may make predictions for Y when, for

78

example, X is increased by 8 units or when X is decreased by 6 units. These predictions rely on simple deductive reasoning:

- “If a one unit increase in X causes a 0.5 unit increase in Y , then increasing X by 8 units will increase Y by 4 units.”
- “If a one unit increase in X causes a 0.5 unit increase in Y , then decreasing X by 6 units will decrease Y by 3 units.”

To add context, suppose X is advertising expenditure (in millions) and Y is sales units (in thousands). Then, our reasoning may be: “If a \$1 million increase in advertising expenditure leads to an increase in sales units of 500, then increasing advertising expenditure by \$8 million will increase sales units by 4,000.”

For these deductive arguments, we are assuming the causal relationship between X and Y , and then making a prediction. Since the conclusion essentially follows from the definition of a causal relationship, disagreement with the conclusion requires disagreement with the assumed causality. Using our advertising example, if you disagree that an \$8 million increase in advertising expenditure will increase sales units by 4,000, then you must disagree with the general (assumed) causal relationship of \$1 million in advertising generating 500 sales units.

To resolve a disagreement about an assumption, we must either show robustness or invoke inductive reasoning. Since our prediction is unlikely to be robust to alternative causal relationships (e.g., \$1 million in advertising generating 900 sales units will lead to a clearly different prediction when

advertising increases by \$8 million), we turn to inductive arguments based on data. Just as we demonstrated for a population mean in this chapter, the process involves a combination of deductive and inductive reasoning to build a confidence interval and/or conduct a hypothesis test for the “true”/population-level causal relationship between X and Y . Specifically, we make a set of assumptions that imply causality between X and Y , and the distribution of an estimator for the magnitude. Then, based on a set of sample statistics we observe in our sample data, we formulate inductive arguments about the population-level relationship between X and Y using a confidence interval and/or hypothesis test.

[Reasoning Box 3.5](#) summarizes this basic paradigm.

REASONING BOX 3.5

REASONING IN ACTIVE PREDICTIONS

The underlying reasoning for active predictions is as follows.

Forming the prediction uses deductive reasoning.

Assume the causal relationship, which then implies the prediction.

Estimating the causal relationship uses deductive and inductive reasoning.

Deductive reasoning: Make assumptions that imply: causality between X and Y and the distribution of an estimator for the magnitude of this causality in the population.

Inductive reasoning: Using an observed data sample, build a confidence interval and/or determine whether to reject a null hypothesis for the magnitude of the population-level causality.

sample mean). What remains is how to establish causality and construct an estimator for the causal relationship. Filling in these pieces is the basis of the next several chapters.

RISING TO THE dataCHALLENGE

Knowing All Your Customers by Observing a Few

Let's return to the Data Challenge posed at the start of the chapter: knowing all your customers by observing a few. The dataset you have is sample data of size $N = 37$. The sample mean is \$41,659, and the sample standard deviation is 4428.35. Therefore, if we assume this is a random sample from the full population of people who have downloaded the app, we can build a confidence interval and conduct a hypothesis test using the reasoning presented in this chapter.

For example, we are 99% confident that the mean income for the entire set of people downloading this app is within approximately 2.58 standard deviations of the sample mean, i.e., between $41,659 - 2.58 \times \left(\frac{4428.35}{\sqrt{37}} \right)$ and $41,659 + 2.58 \times \left(\frac{4428.35}{\sqrt{37}} \right)$. Simplifying, we are 99% confident that the mean income for the entire set of people downloading this app is between \$39,780.72 and \$43,537.28.

Regarding the possibility that the mean income of the entire set of downloaders (i.e., population mean) is \$38,000, we note the following. The observation of a sample mean of \$41,659 with a population mean of \$38,000 generates a t -stat of 5.027. This exceeds 2.58, and so if we stand by our assumption of a random sample, we reject \$38,000 as the population mean with 99% confidence.

Taking an alternative approach, we note that the p -value for this t -stat is 0.000013. Since this is less than 0.01, it is less than 1% likely to observe \$41,659 as our sample mean when the population mean is \$38,000. Again, this

leads to 99% confidence that the mean income of the entire set of downloaders is not \$38,000.

SUMMARY

This chapter introduced random variables and their distributions, as well as basic sample statistics for data samples. It showed how to build confidence intervals and conduct hypothesis tests for population parameters. Throughout the chapter, we illustrated how to incorporate both types of reasoning, deductive and inductive, when drawing conclusions about a general population using a data sample from that population. Lastly, we showed in general terms how to utilize reasoning, along with confidence intervals and/or hypothesis tests, to make active predictions.

The application of basic reasoning to data analysis is a foundation for virtually every topic that follows in this book. The conclusions we will try to draw about general populations will become more intricate, and so the path from sample to population will be more complex. However, the underlying structure will follow very closely the format described here.

80

KEY TERMS AND CONCEPTS

confidence interval

continuous random variable

deterministic variable

discrete random variable

estimator

expected value

hypothesis test

independent (random variable)

independent and identically distributed (i.i.d.)

normal random variable
null hypothesis
p-value
population mean
population parameter
population standard deviation (σ)
population variance
probability density function (pdf)
probability function
random sample
random variable
sample mean
sample of size N
sample standard deviation
sample statistic
sample variance
standard deviation
t-stat
test statistic
unbiased estimator
variance

CONCEPTUAL QUESTIONS connect

1. Consider the following data sample for a variable X: (LO1)

23	15
8	41
32	9

25	48
37	11
29	30
18	36
4	19

- a. Calculate the sample mean for X
 - b. Calculate the sample standard deviation for X
 - c. The sample mean and sample standard deviation are estimators. What are they estimators for?
 - d. Explain what it means to be an unbiased estimator in both formal and informal terms.
2. Suppose a shoe manufacturer randomly selects 1,000 of its customers and asks them their income. Of the 1,000 customers questioned, 472 answer the question and the remainder decline. For those 472, you construct a 99% confidence interval for the income of all of your customers equal to (\$52,817, \$61,247) using the sample mean and standard deviation as described in [Reasoning Box 3.2](#). One of your colleagues challenges your confidence interval, claiming your confidence is too high and the actual mean income is almost certainly below your lower bound. Where is the potential flaw in your reasoning that would generate this critique of your conclusion? (LO2)
3. The click-through rate for an advertisement on a web page is the percentage of visitors to that web page who click the link for the advertisement. Suppose you have sample data on the click-through rate for a website consisting of a random sample of visitors to that website. For each visitor i , you observe C_i , which equals 1 if they clicked on the advertisement and 0 if they did not. Defined this way, \bar{C} is the sample mean and represents the click-through rate for the sample. Suppose the sample size is 2,271, the sample mean is 0.12, and the sample standard deviation is 0.325. (LO3)

-
- a. Calculate a 90% confidence interval for the population mean, i.e., the click-through rate for the entire population of visitors to this website.

- b. Explain what this confidence interval means.
- c. Suppose your sample consists of 17 people instead of 2,271. How does this affect your ability to use a confidence interval?
4. Refer to the information given in Question 3. Now suppose the sample size is 341, the sample mean is 0.07, and the sample standard deviation is 0.256. The advertiser was hoping for a click-through rate of 7.5%. (LO5)
- What is the t -stat if the advertiser's hoped-for rate is correct for the full set of web page visitors?
 - Explain what this t -stat is.
 - What is the p -value for this t -stat?
 - Explain what this p -value is.
5. You have a random sample of 1,627 for the random variable X_i . The sample mean is 26.2 and the sample standard deviation is 39.1. Define μ as the population mean for X_i . Assume that $\mu = 20$. (LO4)
- Sketch the distribution of \bar{X} .
 - Graphically illustrate the p -value associated with $\bar{X} = 26.2$.
6. You have a random sample of 1,627 for the random variable X_i . The sample mean is 26.2 and the sample standard deviation is 39.1. Define μ as the population mean for X_i . (LO2)
- Suppose you have generated a 90% confidence interval and a 95% confidence interval for μ . Without doing the calculation, which confidence interval will be wider (i.e., which confidence interval will have a larger difference between its upper and lower bounds)?
 - Calculate a 100% confidence interval for μ .
7. Refer to Question 5. Now assume that $\mu = 26.2$. (LO5)
- Without using a computer, what is the p -value associated with $\bar{X} = 26.2$?
 - Is there any degree of support for which you would reject the null that $\mu = 26.2$?
8. Your top analyst informs you that he is 95% confident that a \$1 increase in the price of your product will result in somewhere between a \$200,000 and \$215,000 loss in revenue. Using this information, he then predicts that your proposed \$2 increase in price will lower revenues somewhere

between \$400,000 and \$430,000, with his best guess being a decline of \$415,000. Speaking conceptually, where does deductive reasoning play a role in this prediction? (LO6)

QUANTITATIVE PROBLEMS connect

9. You have just opened a restaurant in a large city, and you are deciding what you should charge for a regular-sized soda. You'd like to charge a price equal to the average of your competitors, which you believe is \$2.58. To inform your decision, you want to learn more about the average price charged by competing restaurants in the area. You know you won't be able to get prices for every restaurant, so you randomly sample 35 and collect their soda prices. These data are in *Soda.xlsx*. (LO4)

Dataset available at www.mhhe.com/prince1e

- a. You are assuming the mean soda price is \$2.58 for all of your competitors. When conducting data analysis to test this belief, what is this assumption called?
 - b. Calculate the *t*-statistic assuming the mean soda price for all of your competitors is \$2.58.
 - c. Calculate the *p*-value for your *t*-statistic.
 - d. Using a confidence level of 90%, test whether the mean soda price of all your competitors is \$2.58 using the *t*-stat.
 - e. Using a confidence level of 90%, test whether the mean soda price of all your competitors is \$2.58 using the *p*-value.
 - f. Is it possible that your answers to parts d and e would yield different conclusions?
10. To help make decisions about advertising potential for your website, you are interested in learning the average amount of time visitors to your website spend on the site. You manage to collect a month's worth of data that includes 9,872 website visits and their duration. The data are in *WebVisits.xlsx*. (LO3)

Dataset available at www.mhhe.com/prince1e

- a. Build a 90% confidence interval for the mean visit duration for all visitors to your website. Explain what this confidence interval means.
 - b. Build a 95% confidence interval for the mean visit duration for all visitors to your website. Explain what this confidence interval means.
 - c. Build a 99% confidence interval for the mean visit duration for all visitors to your website. Explain what this confidence interval means.
 - d. Is there reason to believe your confidence levels are inaccurate? If so, what assumption(s) may be inaccurate?
11. To better assess your willingness-to-pay for advertising on others' websites, you want to learn the mean profit per visit for all visits to your website. To accomplish this, you have collected a random sample of 4,738 visits to your website over the past six months. This sample includes information on visit duration and profits. The data are contained in *WebProfits.xlsx*. Using the data in *WebProfits.xlsx*: (LO1)

Dataset available at www.mhhe.com/prince1e

- a. Build a 99% confidence interval for the mean profit per visit for all of your visitors.
 - b. Let the null hypothesis be that mean profit per visit for all of your visitors is \$11.50.
 - i. Calculate the corresponding *t*-stat for this null hypothesis.
 - ii. Calculate the corresponding *p*-value for this null hypothesis.
 - iii. With strength of 95%, decide whether or not to reject this null hypothesis.
 - iv. Detail the reasoning behind your decision.
12. Refer to Problem 11 and use the data in *WebProfits.xlsx*. Let the null hypothesis be that mean profit per visit for all your visitors is \$9.00. (LO6)

Dataset available at www.mhhe.com/prince1e

- a. Calculate the corresponding *t*-stat for this null hypothesis.
- b. Calculate the corresponding *p*-value for this null hypothesis.

- c. With strength of 99%, decide whether or not to reject this null hypothesis.
- d. Detail the reasoning behind your decision.

The Scientific Method: The Gold Standard for Establishing Causality

4

LEARNING OBJECTIVES

After completing this chapter, you will be able to:

- LO4.1** Recall the elements of the scientific method.
- LO4.2** Explain how experiments can be used to measure treatment effects.
- LO4.3** Execute a hypothesis test concerning a treatment effect using experimental data.
- LO4.4** Construct a confidence interval for a treatment effect using experimental data.
- LO4.5** Differentiate experimental from nonexperimental data.
- LO4.6** Explain why using nonexperimental data presents challenges when trying to measure treatment effects.

dataCHALLENGE Does Dancing Yield Dollars?

While you are working as a manager at Chipotle, your store owner wonders aloud whether his practice of hiring college-age adults to dance outside his restaurant with store signs to grab commuters' attention is really worth the investment. Do they really draw more customers to the store, and if so, how many? He then laments that there's no way to really "know," and so deciding whether to keep hiring dancers is just a matter of theory and guess work.

Seeing a golden opportunity to demonstrate your analytical skills, you assure him there is a way to get a good measurement of these dancers' effectiveness. You go on to explain how it can be done.

What approach would you suggest to the owner?

84

Introduction

This chapter describes the scientific method and how it can generate knowledge about causality. As the chapter subtitle indicates, the scientific method is essentially the "gold standard" when it comes to establishing causality with data. Ideally, we would establish all causal relationships using the scientific method, but unfortunately that is often not possible due to data limitations. Nevertheless, it is useful to understand how causality is established under ideal conditions, and then consider issues that arise when these ideal conditions are not met. As you progress through subsequent chapters, it is useful to always have the scientific method and the basic reasoning behind it firmly in mind. Questions like, "How do these data differ from what would be collected using the scientific method?" and "What is the line of reasoning that took us from that sample statistic to concluding there is a causal relationship?" will apply throughout the book and beyond. We conclude this chapter by drawing basic distinctions between data that are generated via the scientific

method and data that are typically available in the business world for analysis.

The Scientific Method

LO 4.1 Recall the elements of the scientific method.

The “scientific method” is often presented and discussed in the context of “hard” sciences, e.g., physics, chemistry, and medicine. It is often used to establish causal relationships between variables. In fact, as we will show shortly, the simple and elegant reasoning structure embedded within the scientific method makes it an ideal means of accomplishing this goal.

Even if you have never seen a formal description of the scientific method, it is likely you are familiar with some of its basic features and applications. (If not, you will be soon!) A classic application is in medicine, where researchers run clinical trials to learn the impact of a new drug on patients’ health outcomes. In such a trial, participants are randomly assigned into two groups, with one group given the new drug and the other given a placebo (e.g., a pill filled with water). Researchers then compare the health outcomes between the two groups to learn about the drug’s impact.

In what follows, we’ll explain how clinical trials like the above drug example, along with some experiments in business, are classic applications of the scientific method. We will also provide a framework that clearly illustrates how and why the scientific method so effectively establishes causality.

DEFINITION AND DETAILS

Put concisely, the **scientific method** is a process designed to generate knowledge through the collection and analysis of experimental data. The full process for the scientific method consists of the following six parts:

scientific method The process designed to generate

knowledge through the collection and analysis of experimental data.

1. Ask a question.
2. Do background research.
3. Formulate a hypothesis.

85

FIGURE 4.1 The Scientific Method Process



4. Conduct an experiment to test the hypothesis.
5. Analyze the data from the experiment and draw conclusions.
6. Communicate the findings.

We describe this process graphically in [Figure 4.1](#).

Let's consider these steps in more detail; in doing so, we will put them in context using examples involving questions of causality in both medicine and business. For our medicine example, we will attempt to measure the effect of a new drug on the incidence of cancer. For our business example, we will

attempt to measure, for a given firm's website, the effect of increasing the size of a banner ad on the click-through rate for the advertiser. A banner ad is an advertisement embedded into a web page that allows visitors to that page to click on it and move to the web page of the advertiser. The click-through rate is the number of times an ad is clicked divided by the number of times it is shown (e.g., if the ad is shown 50 times and clicked once, the click-through rate is $1/50 = 2\%$).

STEP 1. The first step of the scientific method is to ask a question.

Deciding which question to ask can be an organic process, motivated by various observations of the surrounding environment. For example, after observing different objects falling at the same speed, one may ask whether the force of gravity is constant. Often the formulation of a question is quite straightforward, motivated by interest in a particular outcome.

- For our medicine example, the initial question is of the form: “What can reduce the likelihood of cancer?” This question is motivated by our interest in being healthy by avoiding a cancer outcome.
- For our business example, the initial question is of the form: “What can improve the click-through rate for a banner ad?” This question is motivated by our interest in generating more revenue by increasing click-through rates.

STEP 2. The second step of the scientific method (Do background research) involves learning more about the issues surrounding the posed question. The purpose is to find information that will help identify a possible answer to your question, which you will ultimately test.

- For our medicine example, background research may consist of learning about various different chemicals and compounds that have shown some impact on the outbreak of cancer cells.
- For our business example, background research may consist of

consumer surveys aiming to identify what ad features gain their attention.

STEP 3. The third step of the scientific method (Formulate a hypothesis) involves hypothesizing a possible answer to the posed question. In general, a **hypothesis** is a proposed idea based on limited evidence that leads to further investigation. Within the scientific method, the hypothesis is typically grounded in the background research that was done and involves a positive statement about causality (i.e., X causes Y).

hypothesis A proposed idea based on limited evidence that leads to further investigation.

- For our medicine example, background research may lead to the formulation of a new drug believed to reduce the outbreak of cancer cells. The corresponding hypothesis then would be that taking this new drug causes a reduced risk of cancer.
- For our business example, background research may indicate that larger banner ads are more likely to catch the attention of website visitors. This observation may lead to a hypothesis that doubling the size of a banner ad will increase the click-through rate for that ad.

STEP 4. The fourth step of the scientific method involves running an experiment. In the context of the scientific method, an **experiment** is a test within a controlled environment designed to examine the validity of a hypothesis. Data that result from an experiment are called **experimental data**. For hypotheses about causality, the experiment generally involves allocating a binary treatment, or treatment levels, across two or more groups. A **treatment** within an experiment is something that is administered to members of at least one participating group. The treatment is based upon the causal factor (X) in the hypothesis, and the **treatment effect** is the change in the outcome (Y) resulting from variation in the treatment.

experiment A test within a controlled environment designed to examine the validity of a hypothesis.

experimental data Data that result from an experiment.

treatment Something that is administered to members of at least one participating group.

treatment effect The change in the outcome resulting from variation in the treatment.

- For our medicine example, the (binary) treatment might be the drug itself, where one group receives the drug and another group does not. Alternatively, there may be different levels of treatment, where one group receives nothing, another group receives a small drug dose, another a larger dose, and so on.
- For our business example, the treatment would be the doubling of the banner ad size. One group of website visitors would see the normal-sized banner ad (i.e., no treatment), while another group of website visitors would see the double-sized banner ad (i.e., treatment).

87

STEP 5. The fifth step of the scientific method involves analyzing the data from the experiment and drawing conclusions. For the analysis, we compare the measured outcomes between the group receiving the treatment and the group that did not (or across groups with differing treatment levels). In doing so, we typically build a confidence interval for the treatment effect and/or conduct a hypothesis test concerning the size of the treatment effect. Based on such analysis, we draw conclusions about the existence and magnitude of the treatment effect —that is, whether there is a causal relationship and if so, how big it is. Of course, these conclusions will depend on both the numbers resulting from the experiment and a clear line of reasoning.

- For our medicine example, a comparison of cancer rates for those who

took the drug versus those who didn't, along with assumptions and a line of reasoning, may lead us to conclude the drug does reduce the incidence of cancer.

- For our business example, analysis of the click-through rates for the two different banner ad sizes, along with assumptions and a line of reasoning, may lead to the construction of a confidence interval for the effect of doubling the size of a banner ad.

We discuss the full details of this process below.

STEP 6. The final, sixth step in the scientific method is to communicate the findings. This requires the researcher to explain both the methodology and findings. For those conducting experiments with business applications, it is especially important to be able to communicate the statistical analysis and underlying reasoning in a way that a nonexpert will understand. This point holds true both for experimental findings and for analysis involving any of the methods we discuss throughout this book. A full communication of the findings typically consists of a main conclusion, a confidence level, description of the experiment, reasoning leading to the conclusion, and summary of the statistics used.

- For our medicine example, the main conclusion and confidence level may read: "Scientists 99% confident new drug reduces cancer."
- For our business example, they may read: "Researchers 95% confident doubling of banner ad size increases click-through rate between 0.6% and 0.9%."

Both would follow with descriptions of the experiment, relevant reasoning, and the statistics used. We summarize the components of the scientific method for both examples in [Table 4.1](#).

As we have seen, the scientific method generally is used to learn about causal relationships. Why is it so effective at doing this? The intuition for why it works is something you may already have used without realizing it. Suppose

you own a dog, and one day you notice he has developed a rash with no apparent cause. While you plan to take him to the veterinarian to be treated, you would like to learn what caused the rash in order to avoid your dog getting another one in the future. To establish the cause, you may ask yourself a simple question: “What is different?” That is, you want to establish whether anything has changed recently for your dog. Ultimately, if you identify a single, notable change concerning your dog, you are likely to pin that change as the cause for the rash. For example, if you purchased a different dog food the prior week but everything else concerning your dog has remained the same, then you may conclude the new dog food is causing the rash.

88

TABLE 4.1 Summaries of Scientific Method for Medicine and Business Examples

	CANCER DRUG (MEDICINE)	BANNER AD SIZE (BUSINESS)
Question	What can reduce the likelihood of cancer?	What can improve the click-through rate for a banner ad?
Research	Chemical research	Consumer surveys
Hypothesis	New drug reduces risk of cancer.	Doubling a banner ad's size will increase its click-through rate.
Experiment	Give drug to one group and not to another.	Double the banner ad size for some website visitors and not for others.
Analysis/ Conclusion	Compare cancer rates across two groups. Conclude there is an effect.	Compare click-through rates across two groups. Conclude the effect lies in a certain range.
Dissemination	“Scientists 99% confident new drug reduces cancer.” Description of experiment,	“Researchers 95% confident doubling of banner ad size increases click-through rate between 0.6% and 0.9%.” Description of experiment,

reasoning, statistics.

reasoning, statistics.

Experiments used in the scientific method take the above intuition to the extreme. In running an experiment, you do everything you can to ensure there is a single, controlled difference (i.e., treatment) across participating groups. Then, if the treated group's outcome is, say, better than the untreated group's, we feel confident attributing this improvement as the effect of the treatment, since the treatment was the only thing that differed between the two groups. That is the intuition behind how experiments establish causality. Next, we will develop and analyze a framework that helps formalize the reasoning behind why this is the case.

COMMUNICATING DATA 4.1

PENICILLIN AND THE SCIENTIFIC METHOD

One classic example of the scientific method appears in the story of Alexander Fleming's discovery of penicillin from mold spores. Fleming worked at St. Mary's Hospital in London as a bacteriologist. He left on vacation in 1928, and when he returned, he discovered a mold growing in one of the bacteria cultures he'd accidentally left open. The culture was the staphylococcus bacteria, and in the middle of it was growing a strain of mold called *Penicillium notatum*. He observed that the staphylococcus developed throughout the petri dish, except for an area around the mold, which remained clear of bacteria. This observation was where Fleming began the process of refining an antibiotic that would save thousands of lives.

Fleming observed that bacteria would not grow in the area directly surrounding the *Penicillium* mold, and he questioned what was inhibiting the growth. Upon researching the mold, he hypothesized that it secreted something that hampered bacterial growth. To test this hypothesis, Fleming began conducting experiments, using the mold as the treatment and the level of bacterial growth as his outcome of interest. After analyzing the results of the experiments, he concluded that the compound the mold secreted was capable

of killing a variety of bacterial strains.

Upon Fleming's confirmation that the mold would prevent the growth of other bacteria, other scientists began attempting to purify the Penicillium byproduct. The breakthrough came in the 1930s and 1940s during World War II, when scientists finally managed to refine and test the new Penicillin from Oxford University. Afterwards a variety of drug manufacturers began production to supply soldiers and citizens during the war.

THE SCIENTIFIC METHOD AND CAUSAL INFERENCE

LO 4.2 Explain how experiments can be used to measure treatment effects.

In this section, we develop a simple framework to help us understand the difference between what we hope to measure using an experiment, and what the experiment is actually able to measure. We then lay out the basic assumptions and reasoning necessary for us to believe the latter equals the former. Along the way, we will highlight classic reasons why improper experiments can lead to biased conclusions about causality.

A Simple Treatment Framework The basic goal when running an experiment is to measure a treatment effect. As previously defined, the treatment effect is the change in the outcome resulting from variation in the treatment. Put another way, the treatment effect solves the following dilemma: For a given individual, measure his outcome when he does not receive the treatment. Then, for the exact same person, measure his outcome when he does receive the treatment. The difference then is the treatment effect.

Thinking again about clinical trials, consider the treatment effect for a new drug on a person's cholesterol level. Then, for a given individual—say, Mike—we'd like to measure Mike's cholesterol level when he doesn't take

the drug and see what the difference is in the exact-same Mike's cholesterol level when he does take the drug. Of course, this comparison applies beyond just people. Following our banner ad example, we would like to take a single banner ad and compare its click-through rate at its normal size versus the click-through rate for the exact same banner ad at double its size.

We can formalize this idea as follows, using a paradigm sometimes called the potential outcomes framework. Consider a group of subjects who will participate in an experiment. We can index these subjects with the letter i . Thus, $i = 1$ refers to the first subject, $i = 2$ the second, and so on. Now, for any given subject i , consider the outcome realized by that subject if it receives the treatment (T). We denote this as Outcome_i^T . Similarly, consider the outcome realized by that subject if it does not receive the treatment (NT). We denote this as Outcome_i^{NT} . Then, the treatment effect for subject i is simply the difference:

$$\text{Treatment Effect}_i = \text{Outcome}_i^T - \text{Outcome}_i^{NT}$$

Given this characterization of the treatment effect, the process of measuring it for a given subject (person or banner ad) may seem straightforward. We should (1) choose one subject, (2) measure its outcome without the treatment, (3) give it the treatment, (4) measure its outcome with the treatment, and (5) take the difference. Unfortunately, though, such an approach will not reliably measure the treatment effect for that subject. The reason is that when a given subject shifts from not having the treatment to having the treatment, it is no longer the "exact same." For example, during the time between when Mike is untreated (isn't taking the drug) and when he is treated (is taking the drug), he may have changed his diet, fallen ill, began exercising, or some such change. There is no guarantee that the untreated Mike is the "exact same" as the treated Mike. In fact, it is virtually guaranteed that there will be at least some subtle difference in any subject between the time it does not receive the treatment and the time that it does. Consequently, we cannot confidently attribute differences in the cholesterol level between the treated Mike and untreated Mike as the effect of the drug, since other changes besides the

administration of the drug almost certainly occurred during the same time period.

The problem we face in trying to measure a treatment effect is that our subjects cannot be both untreated and treated at the same time. Hence, we must choose a single treatment

90

status at the time of the experiment for any given subject. This means we now need at least two subjects in order to observe an outcome with the treatment and an outcome without the treatment. One subject gets the treatment and the other subject does not get the treatment.

Could we simply take the difference in the outcome between the treated subject and the untreated subject $(\text{Outcome}_i^T - \text{Outcome}_j^{NT})$ and use that as an estimate of the treatment effect? No, because with two subjects involved, there no longer is a single treatment effect. The treatment effect for one subject may be (and often is) different from the treatment effect for another. For example, the effect of a cholesterol drug on one person's cholesterol level may be a reduction of 30 points, while its effect on another person may be a reduction of just 5 points. Taking the difference in outcomes between a treated subject and an untreated subject measures neither subject's treatment effect. To see this, note that:

$$\text{Outcome}_i^T - \text{Outcome}_j^{NT} \neq \text{Outcome}_i^T - \text{Outcome}_i^{NT} = \text{Treatment I}$$

and

$$\text{Outcome}_i^T - \text{Outcome}_j^{NT} \neq \text{Outcome}_j^T - \text{Outcome}_j^{NT} = \text{Treatment I}$$

Since we are unable to measure treatment effects for individual subjects, we instead attempt to estimate the mean, or average, treatment effect (often written as ATE) across the entire population of subjects who may receive the treatment. The **average treatment effect (ATE)** is the average difference in the

treated and untreated outcome across all subjects in a population. In measuring the ATE, we consider each subject's treatment effect as a draw from the population of treatment effects across all possible subjects. Consequently, Treatment Effect_i is a random variable whose distribution mirrors the distribution of treatment effects across the entire population of possible subjects.

average treatment effect (ATE) The average difference in the treated and untreated outcome across all subjects in a population.

To make this more concrete, consider again the cholesterol drug. The effect of the drug will vary across individuals, and for the entire population of potential users of the drug, the effect (measured as number of points reduced in one's cholesterol level) may have a normal distribution with mean of 20 and variance of 15, i.e., Treatment Effect_i ~ $N(20, 15)$. For this example, the ATE would then be 20, and this is the number we would hope to accurately measure using a sample of subjects.

The average treatment effect is simply the expected value of the treatment effect for a randomly drawn subject from the population, written as $E[\text{Treatment Effect}_i]$. Expanding on this, we have:

$$\text{ATE} = E[\text{Treatment Effect}_i] = E\left[\text{Outcome}_i^T - \text{Outcome}_i^{NT}\right]$$

With the estimation of the ATE for a given treatment as our objective, we now turn to how experiments can help up accomplish this goal.

From Experiments to Treatment Effects How can an experiment provide us with an estimate of the average treatment effect? The answer to this question is best understood by considering an experiment with a dichotomous treatment—that is, a treatment in which participants are split into two groups where one receives the treatment (takes a drug) and the other does not (takes a placebo).

To begin, we define two more variables, whose values are determined by the experiment. The first is a dichotomous variable, Treated_i . This variable equals 1 if subject i actually

91

received the treatment during the experiment, and 0 if the subject did not receive the treatment. The second is Outcome_i . This variable equals the outcome actually experienced by subject i after the experiment. With these definitions, note that:

$$\begin{aligned}\text{Outcome}_i &= \text{Outcome}_i^T \text{ if } \text{Treated}_i = 1 \\ \text{Outcome}_i &= \text{Outcome}_i^{NT} \text{ if } \text{Treated}_i = 0\end{aligned}$$

Given this close relationship among the Outcome variables, it may seem unnecessary to create the variable Outcome_i . However, the values for Outcome_i^T and Outcome_i^{NT} do not depend on which group the subject was assigned to (treated group or untreated group). Rather, they are the outcomes that the subject would realize when treated or untreated, respectively, regardless of the group to which the subject was actually assigned. In contrast, Outcome_i does depend on the group assignment; it is the outcome actually realized, which may depend on whether the treatment was actually received or not.

To solidify your understanding of this difference, suppose Mike's cholesterol level falls 21 points if he takes the drug but remains unchanged if he does not take the drug. Further, suppose Mike was given the drug (i.e., he was assigned to the treatment group). Then, $\text{Treated}_i = 1$, $\text{Outcome}_i^T = -21$, $\text{Outcome}_i^{NT} = 0$, and $\text{Outcome}_i = -21$. If, instead, Mike was not given the drug, we then would have $\text{Treated}_i = 0$ and $\text{Outcome}_i = 0$; the other two variables would not change.

When we run an experiment, it will generate sample data consisting of realizations for Outcome_i and Treated_i for all the subjects. Using these data,

we can calculate the mean outcome for those in the treated group ($\overline{\text{Outcome}_i | \text{Treated}_i = 1}$) and the mean outcome for those in the untreated group ($\overline{\text{Outcome}_i | \text{Treated}_i = 0}$). For our drug example, we can calculate the average change in cholesterol level for those receiving the drug, and then the average change in cholesterol level for those who did not receive the drug. If we take the difference in these two measures, it may seem intuitive that it will represent the average treatment effect. In particular, if those taking the drug have an average change in cholesterol levels of -18 and those who don't have an average change in cholesterol levels of -3 , then we may conclude the $\text{ATE} = -18 - (-3) = -15$.

However, would we arrive at such a conclusion if we were told all of the subjects were men? Or, what if the drug were administered only to people weighing over 230 pounds? If either case were true, we may have serious doubt as to whether -15 is a good estimate of the true ATE. The key question then emerges: *When does the difference in the mean outcomes across the treated and untreated groups yield an unbiased estimate of the ATE?* Or, when does the following relationship hold?

$$E \left[\overline{\text{Outcome}_i | \text{Treated}_i = 1} - \overline{\text{Outcome}_i | \text{Treated}_i = 0} \right] = \text{ATE}$$

We propose that this relationship holds for a given experiment when two conditions are satisfied by that experiment:

1. Participants are a random sample of the population.
2. Assignment into the treated group is random.

As we showed in [Chapter 3](#), if participants are a random sample of the population, then means for the data sample are unbiased estimators for their population counterparts. This relationship is true for conditional means as well. Thus, in our case, we have:

$$E \left[\overline{\text{Outcome}_i | \text{Treated}_i = 1} \right] = E[\text{Outcome}_i | \text{Treated}_i = 1]$$

and

$$E \left[\overline{\text{Outcome}_i | \text{Treated}_i = 0} \right] = E [\text{Outcome}_i | \text{Treated}_i = 0].$$

In words, the mean outcomes for those assigned the treatment and those that weren't in the experiment are unbiased estimates of the mean outcomes in the population for those that would have received the treatment and those that wouldn't. Hence, our first condition is straightforward. It ensures that the sample averages we collect are good estimates of the population parameters to which they naturally correspond. It also leads to the following result: If participants are a random sample of the population, then

$$\begin{aligned} E \left[\overline{\text{Outcome}_i | \text{Treated}_i = 1} - \overline{\text{Outcome}_i | \text{Treated}_i = 0} \right] &= \\ E [\text{Outcome}_i | \text{Treated}_i = 1] - E [\text{Outcome}_i | \text{Treated}_i = 0] \end{aligned}$$

That is, the expected difference in the mean outcome for the treated and untreated groups (top part) equals the difference in the mean outcome in the population between those that would have received the treatment and those that wouldn't (bottom part).

Based on our result from condition #1, we now just need condition #2 to ensure that

$E [\text{Outcome}_i | \text{Treated}_i = 1] - E [\text{Outcome}_i | \text{Treated}_i = 0] = \text{ATE}$ in order for the two conditions to imply

$E \left[\overline{\text{Outcome}_i | \text{Treated}_i = 1} - \overline{\text{Outcome}_i | \text{Treated}_i = 0} \right] = \text{ATE}$. In other words, we need random treatment assignment to ensure that the difference in expected outcome between those who are treated and untreated equals the expected difference in outcome when a given individual goes from being untreated to treated (ATE).

To assess whether this is the case, let's consider reasons why the mean

outcome for those who received the treatment might differ from the mean outcome for those who did not receive the treatment. In fact, there are two reasons this might be the case. First, those receiving the treatment may respond to it; that is, there is a non-zero average treatment effect at least for the group who were given the treatment. This non-zero average treatment effect for the group given the treatment is called the **effect of the treatment on the treated (ETT)**. Using our notation, and this definition, we have:

effect of the treatment on the treated (ETT) Average treatment effect for the group given the treatment.

$$\text{ETT} = E \left[\text{Outcome}_i^T - \text{Outcome}_i^{NT} \mid \text{Treated}_i = 1 \right]$$

If an ETT exists, then even if both groups have the same mean outcome when not given the treatment, a difference emerges once the group chosen to get the treatment receives it.

Second, the treated and untreated groups may be starting from different places. In particular, the mean outcome for the treated group may be different from the mean outcome for the untreated group, even if neither actually received the treatment. In such a situation, we say there is a **selection bias** in the experiment. Again using our notation and this definition, we have:

$$\text{Selection Bias} = E \left[\text{Outcome}_i^{NT} \mid \text{Treated}_i = 1 \right] - E \left[\text{Outcome}_i^{NT} \mid \text{Tr} \right]$$

If a selection bias exists, then even if the treated group showed no response to the treatment, their mean outcomes would differ due to differences in their mean outcomes before any treatment was received.

selection bias The mean outcome for the treated group would differ from the mean outcome for the untreated group in the case where neither receives the treatment.

To further illustrate these points, consider again our experiment with the cholesterol drug. Suppose we have 500 subjects who receive the drug and 500 subjects who do not receive the drug. Suppose again that the average

change in cholesterol level for those who took the drug was -18 , and the average change in cholesterol level for those who did not

93

take the drug was -3 . Let's now revisit our two possible reasons for why this difference exists—ETT and selection bias—one at a time.

First, suppose there is no selection bias, meaning both groups would experience an average decline in cholesterol levels of 3 when not given the drug. Then, the difference we observe is due to the drug reducing cholesterol levels for those receiving it by 15 on average. That is, the ETT of -15 explains the difference.

Second, suppose there is no effect of the drug on the cholesterol levels of those who took it—that is, $\text{ETT} = 0$. Then, the difference we observe is due to a difference in how each group's cholesterol level changes when not taking the drug. It must be that those taking the drug would have seen their cholesterol 18 points lower without taking the drug (since $\text{ETT} = 0$). Further, we saw that those who did not take the drug had their cholesterol go down by just 3 points. Hence, the selection bias is -15 .

Of course, the difference between the two groups could also be a combination of the ETT *and* selection bias. For example, the drug may lower cholesterol by 7 for the treated group, and their cholesterol would have gone down by 11 even without the drug. In that case, the ETT is -7 , and selection bias is -8 ($-11 - (-3)$).

We can now express the difference in mean outcomes for treated and untreated subjects as follows:

$$E[\text{Outcome}_i | \text{Treated}_i = 1] - E[\text{Outcome}_i | \text{Treated}_i = 0] = \text{ETT} +$$

Consequently, we must determine whether random treatment assignment implies $\text{ETT} + \text{Selection Bias} = \text{ATE}$. To answer this question, we highlight the key implication of random treatment assignment. If treatment is assigned randomly, then the group to which a participant is assigned will provide *no information* about (1) how he responds to the treatment, or (2) his outcome if

he were not to get the treatment. To make this more concrete, suppose group assignment for our cholesterol example is determined by the flip of a coin, and all those who land heads get the treatment and those who land tails do not. Under such a system, should we expect those who land heads to respond to the treatment differently? Or, should we expect those who land heads to have different untreated cholesterol levels than those who land tails? The answer to both questions is certainly “No.” The result of a coin toss has nothing to do with either of these measures.

Let’s now consider the implication of random treatment assignment for ETT and Selection Bias in turn. To begin, we know $\text{ETT} = E \left[\text{Outcome}_i^T - \text{Outcome}_i^{NT} \mid \text{Treated}_i = 1 \right]$. However, with random treatment assignment, we know assignment to the treatment group provides no information about a subject’s response to the treatment. Hence, conditioning on group assignment is completely uninformative, and so we have:

$$\text{ETT} = E \left[\text{Outcome}_i^T - \text{Outcome}_i^{NT} \mid \text{Treated}_i = 1 \right] = E \left[\text{Outcome}_i^T \right]$$

That is, with random treatment assignment, the effect of the treatment on the treated is equal to the average treatment effect.

Next, recall that:

$$\text{Selection Bias} = E \left[\text{Outcome}_i^{NT} \mid \text{Treated}_i = 1 \right] - E \left[\text{Outcome}_i^{NT} \mid \text{Tr} \right]$$

Again, with random treatment assignment, we know assignment to either group (treated or untreated) provides no information about a subject’s outcome were he not to receive

If experiment participants are a random sample from the population and the treatment is randomly assigned, then the difference in the mean outcomes for the treated group and untreated group is an unbiased estimate of the average treatment effect.

the treatment. Consequently, it is again the case that conditioning on group assignment is completely uninformative, meaning we have:

$$\begin{aligned}\text{Selection Bias} &= E\left[\text{Outcome}_i^{NT} \mid \text{Treated}_i = 1\right] - E\left[\text{Outcome}_i^{NT} \mid \text{Untreated}_i = 1\right] \\ &= E\left[\text{Outcome}_i^{NT}\right] - E\left[\text{Outcome}_i^{NT}\right] = 0\end{aligned}$$

In short, random treatment assignment means there is not selection bias. Putting both of these results together, we have that $\text{ETT} + \text{Selection Bias} = \text{ATE} + 0 = \text{ATE}$.

In [Reasoning Box 4.1](#), we detail the reasoning that leads us from an experiment to ultimately measuring a treatment effect.

4.1

Demonstration Problem

Suppose a grocery store is interested in learning the effect of promoting (via a standing sign) a candy bar available in the checkout line on the incidence of its being purchased. To do so, over the course of two weeks, each hour the store randomly chooses whether to display the sign. It also tracks sales of the candy bar during that time. At the end of the two weeks, the candy bar had been promoted for 168 hours and was not promoted for the other 168 hours. The number of purchases during promotion was 210, and the number of purchases when there was no promotion was 126.

1. What are the “subjects” in this experiment?

2. What is the treatment?
3. What is the relevant outcome for this experiment?
4. Calculate an estimate for the average treatment effect, and explain why you believe it to be unbiased.
5. Suppose the promotion was done only during the hours just preceding traditional lunch and dinner times. Explain why this nonrandom treatment assignment will likely result in:
 - a. $ETT \neq ATE$
 - b. Selection Bias $\neq 0$

Answer:

1. The subjects are each hour during the two-week period; half received the treatment and half did not.
2. The treatment is placing the standing sign in the checkout line.
3. There were 168 hours that received the treatment and 168 hours that did not. The relevant outcome is the number of purchases made in each hour.

95

-
4. Although we don't see the number of purchases for each hour, we need only calculate the average for each group to get the ATE. Consequently, the average purchase per hour for the treated group is $210/168 = 1.25$, and the average purchase per hour for the untreated group is $126/168 = 0.75$. Then, the estimate for $ATE = 1.25 - 0.75 = 0.5$. In short, the promotion appears to raise the average number of purchases in an hour by 0.5. As long as the two weeks chosen weren't unusual for candy sales, this should be an unbiased estimate of ATE, given treatment was randomly assigned.
 5. a. Customers are likely hungry during the times right before lunch and dinner, so they may be more prone to respond to a suggestion to purchase a candy bar compared to other times when they are less hungry.

- b. Even without a suggestion to buy a candy bar, hungry customers may be more likely to purchase a candy bar in general than when they are less hungry.

DATA ANALYSIS USING THE SCIENTIFIC METHOD

In this section, we will present an example of the scientific method in practice, detailing the accompanying data analysis. In particular, we will describe the reasoning and analysis that lead to confidence intervals and hypothesis tests for the treatment effect.

Suppose a major search engine company, SearchIt, also produces and sells tablets. The company is currently trying to gain critical market share in the tablet market and is considering various options to make that happen. Its initial question is, “What simple strategies can our company employ to gain more attention and ultimately market share for our tablets?”

COMMUNICATING DATA 4.2

THE EFFECT OF BANNER AD FEATURES

What features of a banner ad get website visitors to click it? Given the immense size of the market for web advertising, answers to this question can be quite valuable. Certainly many firms have conducted their own internal, proprietary research toward this endeavor, but there is also a substantial amount of academic research on the same topic. For example, Lees and Healey (2005)^{*} sought to measure whether adding the image of a mouse clicker arrow next to a “Click here” statement at the bottom of an ad increased the click-through rate for that ad. They conducted their analysis by following the principles of the scientific method. In particular, they took a given ad and made an identical version with the mouse-clicker arrow added. Then, for a

group of websites, they randomly determined which ad each visitor would see. Hence, the group seeing the original ad is the untreated group, and the group seeing the altered ad (with the arrow) is the treated group. They conducted this (field) experiment over a four-week period for three websites.

Knowing this information, what does any difference in click-through rate between the treated and untreated groups tell us? This experiment clearly satisfies the requirement of a random treatment; therefore, we just need to ascertain whether it is a random sample from the population. If the population is “visitors to the three websites used in the experiment,” then as long as there is nothing unusual about the four weeks observed, the difference between the two groups is a reasonable measure of the ATE for the population, i.e., visitors of those websites. If the population is “visitors to any website,” then it is less reasonable to treat the difference in click-through rates between the two groups as the ATE of the population. This is because the three websites chosen are unlikely a random sample of websites (making their visitors a nonrandom sample of website visitors), and even if it were, the sample of websites is too small to represent all websites.

* Lees, G. & Healey, B. “A Test of the Effectiveness of a Mouse Pointer Image in Increasing Click through for a Web Banner Advertisement,” Marketing Bulletin, 2005, 16, Research Note 1.

TABLE 4.2 Sample of Position and Click-through Data for SearchIt Tablet Ad

SEARCH	POSITION	AD CLICKED
1	Top	1
2	Fourth	0
3	Top	0
4	Top	0

5	Fourth	1
6	Fourth	0
...

While researching various options, SearchIt's research team notes that advertising through its own search engine may increase visits to its tablet website and, ultimately, sales. Currently, SearchIt sells the top four search results for the word "tablet" via an auction to any company wanting to advertise its tablets. However, SearchIt is able to reserve one of its advertised slots to promote its own tablet. Companies are typically willing to pay more to have a higher position on the search results, so reserving the top spot for its tablet generally will be more costly than reserving, say, the fourth spot. In deciding which advertising spot to reserve for its own tablet, the company must weigh this cost against the benefit in the form of higher click-through rate of having a higher spot on the search results. Consequently, the hypothesis the company would like to test is: *Higher positioning among the advertised search results leads to higher click-through rates.* In addition, if evidence of such an effect is found, SearchIt would like to be able to quantify it.

To gain knowledge about this hypothesis, SearchIt decides to capitalize on its unique position spanning the search and tablet markets by conducting an experiment. For its next 100,000 searches for the word "tablet," it randomly alters the placement of the ad for its tablet between the top spot and the fourth position among its ad results. For each of these searches, the company records the positioning of its ad and whether the searcher clicked on it. The data may look as shown in [Table 4.2](#).

With data from the experiment in hand, SearchIt's researchers want to analyze these data, arrive at appropriate conclusions, and communicate their findings to their managers. To do so, they conduct both a hypothesis test and build a confidence interval for the impact of ad positioning. We describe both next.

Hypothesis Testing for the Treatment Effect When attempting to measure

the effect of one variable on another, it is common practice to first establish that an effect exists (via a hypothesis test), and then determine a reasonable range for its actual magnitude (via a confidence interval). In this section, we describe how to establish an effect exists using a hypothesis test.

LO 4.3 Execute a hypothesis test concerning a treatment effect using experimental data.

As we have seen in [Chapter 3](#), the process of conducting a hypothesis test requires us to state a null hypothesis to be tested. We'd like to determine whether an effect exists, so it may seem obvious that the null should be something like: A change in ad position from fourth to first affects click-through rates. However, when we use data to inductively reason whether or not the null is true, we can only make strong statements when the data cause us to reject the null—e.g., “We reject the null with 95% confidence.” In contrast, if we instead fail to reject the null, we merely fail to find evidence against it, which is only weak support of it.

We can make a much stronger inductive argument that there is an effect of ad positioning by rejecting the claim that there is not an effect, as opposed to failing to reject that there

97

is one. Given this asymmetry, to be able to make a strong statement about there being an impact of ad placement on click-through rates, our null hypothesis should be:

H_0 : Changing an ad's placement from fourth to first position has (on average) no impact on click-through rates.

If we define the incidence of a click (= 0 or 1) as the outcome and the movement of an ad from fourth to first position as the treatment, then we have the standard null hypothesis when testing for a treatment effect, namely, that the average treatment effect is zero:

$$H_0 : \text{ATE} = E[\text{Outcome}_i^T - \text{Outcome}_i^{NT}] = 0$$

The data from our experiment will provide us with the average incidence of a click (the click-through rate) for ads in top position and the click-through rate for ads in fourth position. Hence, the data provide us with $\overline{\text{Outcome}_i | \text{Treated}_i = 1}$ and $\overline{\text{Outcome}_i | \text{Treated}_i = 0}$.

Following [Reasoning Box 3.1](#), we know that, for a large, random sample (as we have here), these sample means have the following distributions:

$$\overline{\text{Outcome}_i | \text{Treated}_i = 1} \sim N\left(\mu_1, \frac{\sigma_1}{\sqrt{N_1}}\right)$$

$$\overline{\text{Outcome}_i | \text{Treated}_i = 0} \sim N\left(\mu_0, \frac{\sigma_0}{\sqrt{N_0}}\right)$$

Here we define: $\mu_1 = E[\text{Outcome}_i | \text{Treated}_i = 1]$, $\mu_0 = E[\text{Outcome}_i | \text{Treated}_i = 0]$, $\sigma_1 = \sqrt{\text{Var}[\text{Outcome}_i | \text{Treated}_i = 1]}$, $\sigma_0 = \sqrt{\text{Var}[\text{Outcome}_i | \text{Treated}_i = 0]}$, N_1 = the number of treated observations, and N_0 = the number of untreated observations.

A well-known property in statistics is that the sum, or difference, of normal random variables is also a normal random variable—adding or subtracting does not change the shape of the distribution. As a result, we know that the difference in the click-through rates between the treated and untreated observations is normally distributed, since each click-through rate is normally distributed. Using this fact along with basic formulas for expected value and variance, we now have:

$$\overline{\text{Outcome}_i | \text{Treated}_i = 1} - \overline{\text{Outcome}_i | \text{Treated}_i = 0} \sim N\left(\mu_1 - \mu_0, \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_0^2}{N_0}}\right)$$

Lastly, as we proved earlier in this chapter, we know that random treatment assignment implies

$E[\text{Outcome}_i | \text{Treated}_i = 1] - E[\text{Outcome}_i | \text{Treated}_i = 0] = \text{ATE}$. Thus, given our definitions for μ_1 and μ_0 , we have $\mu_1 - \mu_0 = \text{ATE}$ and

$$\overline{\text{Outcome}_i | \text{Treated}_i = 1} - \overline{\text{Outcome}_i | \text{Treated}_i = 0} \sim N\left(\text{ATE}, \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_0^2}{N_0}}\right)$$

Following [Reasoning Box 4.2](#), if we add our null hypothesis that $\text{ATE} = 0$ to our set of assumptions, this means we now have that:

$$\overline{\text{Outcome}_i | \text{Treated}_i = 1} - \overline{\text{Outcome}_i | \text{Treated}_i = 0} \sim N\left(0, \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_0^2}{N_0}}\right)$$

98

REASONING BOX 4.2

THE DISTRIBUTION OF EXPERIMENTAL OUTCOMES

For a given experiment with N participants and a single, binary treatment:

IF:

1. The set of participants is a random sample from the population.
2. The number of subjects (N) is large, such that there are more than 30 in the treated and untreated groups.
3. The assignment of the treatment is random.

THEN:

The difference in the average outcome for the treated group and the average outcome for the untreated group is normally distributed with a mean equal to

the ATE and standard deviation of $\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_0^2}{N_0}}$. In short, we have:

$$\overline{\text{Outcome}_i \mid \text{Treated}_i = 1} - \overline{\text{Outcome}_i \mid \text{Treated}_i = 0} \sim N \left(\text{ATE}, \sqrt{\frac{s_1^2}{N_1} + \frac{s_0^2}{N_0}} \right)$$

Notice that [Reasoning Box 4.2](#), along with our null that the ATE is zero, leads us to an empirically testable conclusion. As described in [Chapter 3](#), we can now calculate our test statistic, which we ultimately will use to assess whether our null hypothesis of the ATE being zero seems credible. Recall that our test statistic simply measures, for a single draw of a random variable, the number of standard deviations that draw is from the mean. Here, our test statistic then is:

$$t = \left(\frac{\overline{\text{Outcome}_i \mid \text{Treated}_i = 1} - \overline{\text{Outcome}_i \mid \text{Treated}_i = 0} - 0}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_0^2}{N_0}}} \right)$$

where we have replaced the population standard deviations (σ) with sample standard deviations (S). Then, using this value directly, we can compare it to 1.65, 1.96, or 2.58 in order to reject or fail to reject the null hypothesis, depending on whether we choose a confidence level of 90%, 95%, or 99%, respectively. Alternatively, we may calculate the p -value of our test statistic and compare this to 10%, 5%, or 1%, depending on our desired confidence level (90%, 95%, or 99%, respectively).

To see how this works for SearchIt, suppose the company's experiment produced the sample statistics reported in [Table 4.3](#).

Using these numbers, we have:

$$N_1 = 49,872; N_0 = 50,128$$

$$\overline{\text{Outcome}_i \mid \text{Treated}_i = 1} = 0.0782; \overline{\text{Outcome}_i \mid \text{Treated}_i = 0} = 0.07$$

$$S_1 = 0.268486; S_0 = 0.259173$$

TABLE 4.3 Sample Statistics for SearchIT

YEAR	PROFITS (MILLIONS)
Number of searches with ad in top position	49,872
Number of searches with ad in fourth position	50,128
Click-through rate for top position ad	0.0782
Click-through rate for fourth position ad	0.072415
Standard deviation of clicks for ad in top position	0.268486
Standard deviation of clicks for ad in fourth position	0.259173

Plugging these numbers into our test statistic gives us:

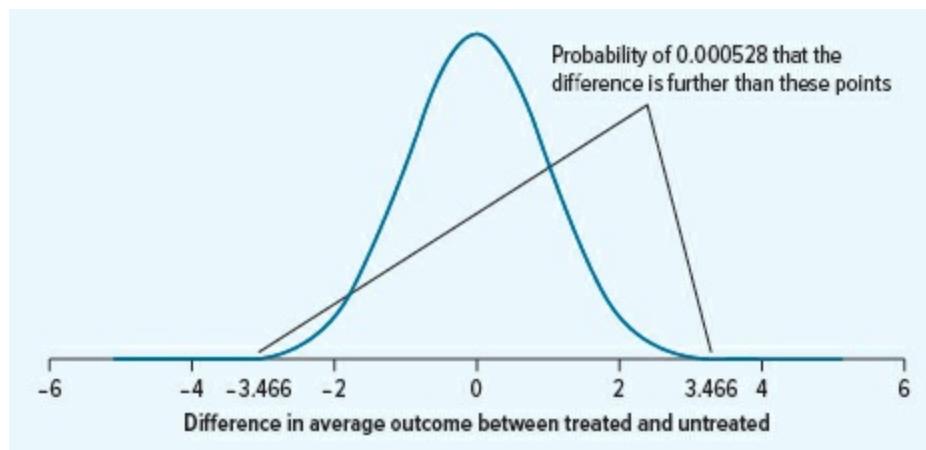
$$t = \left(\frac{(0.0782 - 0.072415) - 0}{\sqrt{\frac{(0.268486)^2}{49,872} + \frac{(0.259173)^2}{50,128}}} \right) = 3.466$$

This test statistic exceeds our cutoff of 2.58, and so we reject the null hypothesis of a zero average treatment effect with 99% confidence.

Alternatively, we can calculate the p -value of this test statistic (using $2 \times (1 - \text{norm.s.dist}(3.466, \text{true}))$ in Excel). This p -value is approximately 0.000528, which is extremely close to zero, and easily less than 1%, again leading us to reject the null of $\text{ATE} = 0$. We illustrate this result in [Figure 4.2](#).

We summarize the general reasoning behind a hypothesis test for a treatment effect in [Reasoning Box 4.3](#).

FIGURE 4.2 P -value for T -stat of 3.466



Confidence Interval for the Treatment Effect Using a hypothesis test, we are able to find evidence against, or consistent with, a specific value of the average treatment effect. As detailed above, the natural choice for a specific value to test is zero, since evidence against will establish that the treatment does have an effect. After accomplishing this, we often want to determine the range of plausible values for the average treatment effect. We do this by using a confidence interval with an objective degree of support.

LO 4.4 Construct a confidence interval for a treatment effect using experimental data.

100

REASONING BOX 4.3

HYPOTHESIS TEST FOR THE TREATMENT EFFECT

Deductive reasoning:

For a given experiment with N participants and a single, binary treatment:

IF:

1. The set of participants is a random sample from the population.
2. The sample size N is large, so that there are at least 30 participants in the

treated and untreated groups.

3. Assignment of the treatment is random.
4. The average treatment effect is zero ($ATE = 0$).

THEN:

The difference in the average outcome for the treated and untreated groups is distributed as:

$$\overline{\text{Outcome}_i \mid \text{Treated}_i = 1} - \overline{\text{Outcome}_i \mid \text{Treated}_i = 0} \sim N \left(0, \sqrt{\frac{\sigma_1^2}{N_1}} + \right)$$

. This difference will fall within 1.65 (1.96, 2.58) standard deviations of 0 approximately 90% (95%, 99%) of the time.

Inductive reasoning:

Using t-stats. If the absolute value of the t -stat is greater than 1.65 (1.96, 2.58), reject the deduced (above) distribution for the difference in sample means. Otherwise, fail to reject. The objective degree of support for this inductive argument is 90% (95%, 99%).

Using p-values. If the p -value of the t -stat is less than 0.10 (0.05, 0.01), reject the deduced (above) distribution for the difference in the sample means. Otherwise, fail to reject. The objective degree of support for this inductive argument is 90% (95%, 99%).

Transposition:

If inductive reasoning leads to a rejection of the distribution for the difference in sample means, reject at least one of the assumptions (1, 2, 3, or 4 above) leading to that distribution. If the sample is large, and there is confidence in a random sample and random treatment assignment, this means rejection of the null hypothesis.

From [Reasoning Box 4.2](#), we know that with a large, random sample of participants and random treatment assignment, we have:

$$\overline{\text{Outcome}_i | \text{Treated}_i = 1} - \overline{\text{Outcome}_i | \text{Treated}_i = 0} \sim N \left(\text{ATE}, \sqrt{\right.$$

From [Chapter 3](#), we know that a normal random variable falls within 1.65 (1.96, 2.58) standard deviations of its mean approximately 90% (95%, 99%) of the time. Applying this to the difference in mean outcomes for the treated and untreated, we have:

$$\Pr \left(\overline{\text{Outcome}_i | \text{Treated}_i = 1} - \overline{\text{Outcome}_i | \text{Treated}_i = 0} \in \left[\text{ATE} \pm 1.65 \times \text{SE} \right] \right) \approx 0.90$$

$$\Pr \left(\overline{\text{Outcome}_i | \text{Treated}_i = 1} - \overline{\text{Outcome}_i | \text{Treated}_i = 0} \in \left[\text{ATE} \pm 1.96 \times \text{SE} \right] \right) \approx 0.95$$

101

4.2

Demonstration Problem

Suppose a yogurt manufacturer is attempting to learn the effectiveness of its latest television advertisement on its sales. It decides to run an experiment where it randomly selects 108 of its 216 markets in which it will run the ad locally. It then records sales for each market. After the experiment is completed, the manufacturer has collected the following sample statistics:

Average sales in markets with the ad: 214,191

Average sales in markets without the ad: 211,382

Standard deviation of sales in markets with the ad: 32,852

Standard deviation of sales in markets without the ad: 31,739

Is there evidence that the ad was effective?

Answer:

In order to get convincing evidence that there is an effect, we must be able to strongly reject the hypothesis that the ad has no effect. Thus, our null hypothesis is: $H_0: \text{ATE} = 0$. To test this, we must construct our t -stat: This is the number of standard deviations the observed difference in average sales lies from the mean of zero. Hence, the t -stat is:

$$t = \left(\frac{(214191 - 211382) - 0}{\sqrt{\frac{(32852)^2}{108} + \frac{(31739)^2}{108}}} \right) = 0.639$$

This t -stat is less than even our lowest cutoff of 1.65. Hence, we would fail to reject the hypothesis that the ad was ineffective in bolstering sales.

We arrive at the same conclusion if we instead use p -values. Here, we have that the p -value for our t -stat is 0.523. This is more than 0.10, so we again fail to reject that the ad was ineffective.

Overall, despite finding higher average sales in markets with the ad, our analysis shows that this difference could very easily have occurred if the ad had no real impact.

$$\Pr \left(\overline{\text{Outcome}_i | \text{Treated}_i = 1} - \overline{\text{Outcome}_i | \text{Treated}_i = 0} \in \left[\text{ATE} \pm \right. \right.$$

In [Figure 4.3](#), we illustrate this idea for the case of 1.96 standard deviations when $\text{ATE} = 0$.

As we did in [Chapter 3](#), we can use some algebra, along with the fact that replacing the population standard deviation (σ) with the sample standard deviation (S) has little impact for large samples, to arrive at the following:

$$\Pr \left(\text{ATE} \in \left[\overline{\text{Outcome}_i | \text{Treated}_i = 1} - \overline{\text{Outcome}_i | \text{Treated}_i = 0} \right] \right)$$

$$\Pr \left(\text{ATE} \in \left[\overline{\text{Outcome}_i | \text{Treated}_i = 1} - \overline{\text{Outcome}_i | \text{Treated}_i = 0} \right] \right)$$

102

COMMUNICATING DATA 4.3

MUSIC TRAINING AND INTELLIGENCE

A recent study by Samuel Mehr attempted to assess whether music education affects an individual's overall intelligence. In his study, he gave a group of subjects a test of general intelligence and then divided the group into two subgroups. The first received music training, and the second received visual arts training. He then tested both groups' general intelligence again after each received their training.

Suppose the study had 100 participants in total, and each subgroup had 50 participants. Suppose also that the general intelligence test was graded on a scale of 0–100. If the “music training” group saw a mean change in its intelligence score of 5 (with standard deviation of 11.3), and the “visual arts” group saw a mean change in its intelligence score of 3 (with standard deviation of 10.7), then Mehr can run a simple *t*-test to determine whether music training made a significant difference, relative to visual arts. Here, the *t*-stat would be:

$$\frac{5 - 3 - 0}{\sqrt{\frac{11.3^2}{50} + \frac{10.7^2}{50}}} = 0.909$$

From this *t*-stat, we fail to reject the claim that both groups had the same change in intelligence.

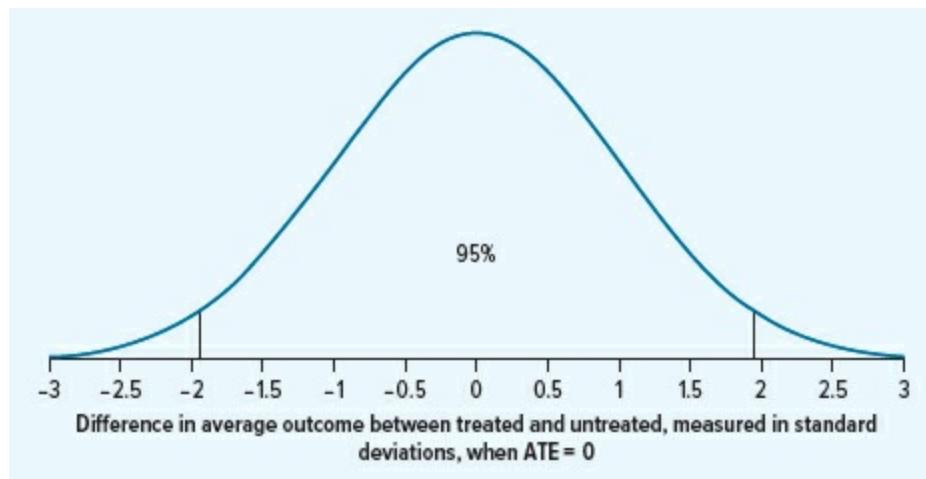
The numbers we used for this study were hypothetical, but the design as described is the design used by Mehr. It is useful to note that, with this design, the analysis is relevant for the effect of music training relative to visual arts

training; there is no true placebo/untreated group. Consequently, our finding of no difference could be due to music training having no effect, but it also could have been the case that both types of training improved intelligence by a comparable amount. In this latter scenario, there is an effect of music training, but it's not distinguishable from the effect of visual arts training.

$$\Pr \left(\text{ATE} \in \left[\overline{\text{Outcome}_i | \text{Treated}_i = 1} - \overline{\text{Outcome}_i | \text{Treated}_i = 0} \pm \right. \right.$$

In words, we can take the difference between the mean outcome for the treated and untreated, and be 95% confident that the true average treatment effect is somewhere within 1.96 standard deviations of that number.

FIGURE 4.3 95% Confidence Interval When ATE = 0



Knowing this, let's build confidence intervals for our SearchIt example. Using the figures from [Table 4.3](#) again, we have:

$$N_1 = 49,872; N_0 = 50,128$$

$$\overline{\text{Outcome}_i \mid \text{Treated}_i = 1} = 0.0782; \overline{\text{Outcome}_i \mid \text{Treated}_i = 0} = 0.07$$

$$S_1 = 0.268486; S_0 = 0.259173$$

Using these numbers, we can build the following confidence intervals for the average treatment effect:

90% confidence interval: (0.003031, 0.008539)

95% confidence interval: (0.002514, 0.009056)

99% confidence interval: (0.001479, 0.010091)

We can then interpret the 99% confidence interval as follows: We are 99% confident that moving an ad from the fourth position to first position will, on average, increase the click-through rate by somewhere between 0.15 and 1.01 percentage points (rounding to two decimal places).

We summarize the reasoning behind building a confidence interval for a treatment effect in [Reasoning Box 4.4](#).

REASONING BOX 4.4

CONFIDENCE INTERVAL FOR THE TREATMENT EFFECT

Deductive reasoning:

IF:

1. The set of participants are a random sample from the population.
2. The sample size N is large, so that there are at least 30 participants in the treated and untreated groups.
3. Assignment of the treatment is random.

THEN:

The interval consisting of the difference between the average outcome for the treated and the untreated, plus or minus 1.65 (1.96, 2.58) standard

deviations for this difference, will contain the average treatment effect approximately 90% (95%, 99%) of the time.

Inductive reasoning:

We observe the difference between the average outcome for the treated and untreated ($\overline{\text{Outcome}_i \mid \text{Treated}_i = 1} - \overline{\text{Outcome}_i \mid \text{Treated}_i = 0}$), the sample standard deviations for the treated (S_1) and untreated (S_0), and the number of subjects receiving the treatment (N_1) and not receiving the treatment (N_0). We conclude the ATE is contained in the interval

$$\overline{\text{Outcome}_i \mid \text{Treated}_i = 1} - \overline{\text{Outcome}_i \mid \text{Treated}_i = 0} \pm 1.65 \left(\sqrt{\frac{S_1^2}{N_1}} + \right)$$

The objective degree of support for this inductive argument is 90%. If we instead use the intervals

$$\overline{\text{Outcome}_i \mid \text{Treated}_i = 1} - \overline{\text{Outcome}_i \mid \text{Treated}_i = 0} \pm 1.96 \left(\sqrt{\frac{S_1^2}{N_1}} + \right)$$

and

$$\overline{\text{Outcome}_i \mid \text{Treated}_i = 1} - \overline{\text{Outcome}_i \mid \text{Treated}_i = 0} \pm 2.58 \left(\sqrt{\frac{S_1^2}{N_1}} + \right)$$

, the objective degree of support becomes 95% and 99%, respectively.

4.3

Demonstration Problem

Consider again the yogurt manufacturer from [Demonstration Problem 4.2](#). Suppose after seeing the disappointing results for its television ad, the firm decides to hire a new advertising team that ultimately produces a new ad. To

determine the effectiveness of the new ad, the manufacturer decides to run another experiment. Like before, it randomly selects 108 of its 216 markets in which it will run the new ad locally. It then records sales for each market. After this new experiment is completed, the manufacturer has collected the following sample statistics:

Average sales in markets with the new ad: 288,307

Average sales in markets without the new ad: 205,191

Standard deviation of sales in markets with the new ad: 34,518

Standard deviation of sales in markets without the new ad: 31,433

Is there evidence that the ad was effective?

Provide a plausible range for the effectiveness of the ad, and provide a corresponding confidence level.

Answer:

As in [Demonstration Problem 4.2](#), in order to get convincing evidence that there is an effect, we must be able to strongly reject the hypothesis that the ad has no effect. Thus, our null hypothesis is: $H_0: \text{ATE} = 0$. To test this, we construct our t -stat:

$$t = \left(\frac{(288307 - 205191) - 0}{\sqrt{\frac{(34518)^2}{108} + \frac{(31433)^2}{108}}} \right) = 18.502$$

This number far exceeds our highest cutoff of 2.58. Hence, we reject the hypothesis that the ad was ineffective, and do so with 99% confidence.

If we instead use p -values, we see that the p -value for our t -stat is essentially zero. This is clearly less than 0.01, so we again reject that the ad was ineffective with 99% confidence.

Now that we have established the ad was effective, we can determine a range of values for the level of its effectiveness. Suppose we want to build a range in which we have 95% confidence. Then, following [Reasoning Box 4.3](#), the range must be:

$$(288, 307 - 205, 191) \pm 1.96 \left(\sqrt{\frac{(34518)^2}{108} + \frac{(31433)^2}{108}} \right) = (74, 311, 91, 921)$$

Summarizing, we are 95% confident that the ad increased sales by somewhere between 74,311 and 91,921.

Experimental Data vs. Non-Experimental Data

As we've highlighted throughout this chapter, experimental data are well-suited toward measuring causal effects of treatments. However, most data that are available to businesses are **nonexperimental data**—data that were not produced using an experiment.

nonexperimental data Data that were not produced using an experiment.

When data are produced outside of an experimental setting, we are no longer able to control how the treatment is administered. Consequently, the treatment is very seldom

105

COMMUNICATING DATA 4.4

MARSHMALLOWS AND RELIABILITY

In a recent study, researchers Celeste Kidd, Holly Palmeri, and Richard Aslin sought to determine whether the reliability of one's environment substantially affected a child's willingness to delay gratification. To do so, they revisited the famous marshmallow test, where children are given one marshmallow but promised a second marshmallow if they can refrain from eating the first until the second arrives.

Kidd et al. altered this test by dividing a group of children into a group facing a reliable environment and a group facing an unreliable environment.

Here, the reliable environment was established by promising other desirable things (e.g., crayons, stickers) and delivering; the unreliable environment was established by promising these same things but not delivering. Once (un)reliability was established, the researchers conducted the marshmallow test—giving the child a marshmallow but promising a second will arrive if she can hold off on eating the first.

For this experiment, the treatment is the difference in reliability, and the outcome is the amount of time the child waits before losing patience and eating the marshmallow. Suppose we ran this experiment with 80 children, randomly assigning them into the reliable and unreliable environments. Suppose also that the mean waiting time for the unreliable group was 3 minutes (with standard deviation of 4.6), and the mean waiting time for the reliable group was 12 minutes (with standard deviation of 6.1). Then, our 95% confidence interval for the difference between the unreliable and reliable groups would be

$$(12 - 3) \pm 1.96 \times \left(\sqrt{\frac{4.6^2}{40} + \frac{6.1^2}{40}} \right) = (6.63, 11.37).$$

Consequently, we would be 95% confident that the difference in time a child is willing to wait for an extra marshmallow before consuming one she already has in a reliable environment vs. an unreliable environment is 6.63 to 11.37 minutes.

randomly assigned, which can confound our ability to properly estimate a treatment effect. For example, if we view price as the treatment and sales as the outcome, it is almost certainly the case that the treatment was *not* randomly assigned across markets. In this section, we discuss some common examples of nonexperimental data that are used in business, and then discuss the consequences of using such data to estimate treatment effects without accounting for their nonexperimental features.

EXAMPLES OF NONEXPERIMENTAL DATA IN BUSINESS

LO 4.5 Differentiate experimental from nonexperimental data.

In business, there are many cases where we want to know the effect of a treatment but have access only to nonexperimental data in order to measure it. [Table 4.4](#) presents a few examples.

All of the examples in [Table 4.4](#) involve a strategic variable a firm may choose to alter (i.e., treatment) and an outcome it believes may be affected by the corresponding treatment.

Now, suppose a firm collects data on these treatment/outcome combinations in order to identify the effect(s) of the treatment(s). For example, a firm may wish to know how a change in price will affect its sales, and it proceeds to collect data on prices and sales across all regions where it has operated for the past year. To simplify the exposition, suppose the firm price-discriminates (that is, it charges different prices) across 32 regions and may change its prices on a monthly basis. To further simplify, suppose the firm chooses among just two price options: a “regular” price of \$25 and a “sale” price of \$15. In this simplified scenario,

106

TABLE 4.4 Examples of Nonexperimental Business Treatments and Outcomes

TREATMENT	OUTCOME
Price	Profits
Price	Sales
Salary	Longevity
Advertising expenditure	Sales
Search engine placement	Website hits
Quality investment	Customer complaints
Employee training	Productivity

we have a single treatment, which is the price decrease from \$25 to \$15, and

the firm wishes to know the effect of this treatment on sales. The firm collects data over the past 12 months across all 32 different regions. [Table 4.5](#) contains a subset of these panel data.

If these were experimental data to be used to measure a treatment effect, we would have randomly varied price across regions and time. However, these are nonexperimental data, and there are plenty of reasons to believe that they were generated in a way where price was not randomly assigned. In fact, we should fully expect that they were set in a very nonrandom way. Prices are generally set by a management team that is trained to set prices in an optimal way according to the market conditions they observe. For example, if Region A has a customer base that is largely wealthier than Region B, management may assume customers in Region A are less price-sensitive and may consequently charge higher prices there relative to Region B. This assumption creates a correlation between price levels and customer wealth, which means we do not have random treatment assignment.

TABLE 4.5 Panel Data on Price and Sales

REGION	MONTH	PRICE	SALES
1	1	15	212
1	2	25	187
1	3	15	230
1	4	15	192
1	5	25	201
1	6	25	172
1	7	15	251
1	8	15	233
1	9	15	195
1	10	25	180
1	11	25	207
1	12	15	219

2	1	25	332
2	2	15	351

CONSEQUENCES OF USING NONEXPERIMENTAL DATA TO ESTIMATE TREATMENT EFFECTS

LO 4.6 Explain why using nonexperimental data presents challenges when trying to measure treatment effects.

With nonexperimental data, there is a high likelihood that the treatment is not randomly assigned. Earlier in this chapter, we showed that:

$$E[\text{Outcome}_i | \text{Treated}_i = 1] - E[\text{Outcome}_i | \text{Treated}_i = 0] = \text{ETT} +$$

This means that, by finding the difference in the mean outcome for the treated and untreated in our random sample, we get an estimator for ETT + Selection Bias. Further, we know random treatment assignment ensures that the effect of the treatment on the treated (ETT) equals the average treatment effect (ATE), and the selection bias equals zero. Thus, the difference in the mean outcomes between the treated and untreated serves as an estimator for the ATE. If treatment assignment is nonrandom, then we risk the possibility that $\text{ETT} \neq \text{ATE}$, $\text{Selection Bias} \neq 0$, or both. If this is the case, comparing the means between the treated and untreated groups is no longer a proper estimator for the ATE.

Consider again our pricing example. Here, price assignment is nonrandom in that it is correlated with the wealth of the regions. In particular, we tend to see higher prices in wealthier regions. How can this nonrandom assignment of price cause $\text{ETT} \neq \text{ATE}$ or $\text{Selection Bias} \neq 0$? First, suppose wealthier customers are, in fact, less price-sensitive than customers who are poorer. Then, a price decrease from \$25 to \$15 will have a larger impact on

sales for poorer regions. Further, poorer regions experienced the lower price of \$15 more often than wealthy regions. Consequently, the effect of the treatment (price decrease from \$25 to \$15) on the treated (consisting mostly of poorer regions) is not the same as the average treatment effect ($\text{ETT} \neq \text{ATE}$). In fact, the ETT is likely more than the ATE.

Second, suppose wealthier customers, all else equal, buy more of the product. This means that, for a given price, we will see higher sales in a market with wealthy customers than one with customers who are less wealthy. Consequently, when the price is \$25, we expect to have higher sales in wealthy regions. Hence, on average, the treated group (consisting mostly of poorer regions) would have lower sales than the untreated group (consisting mostly of wealthy regions) when neither receives the treatment (i.e., each has price of \$25). This conclusion implies there is a non-zero selection bias. In fact, it suggests we have a negative selection bias.

To summarize, the fact that higher prices tended to occur in regions with wealthier customers resulted in both $\text{ETT} > \text{ATE}$ and $\text{Selection Bias} < 0$. Consequently, we have no way of knowing whether $\text{ETT} + \text{Selection Bias} = \text{ATE}$ anymore; we could have $\text{ETT} + \text{Selection Bias} > \text{ATE}$, $\text{ETT} + \text{Selection Bias} < \text{ATE}$, or $\text{ETT} + \text{Selection Bias} = \text{ATE}$. The bottom line: Comparing the mean outcomes for the treated and untreated no longer gives us a reliable estimate of the average treatment effect, due to the nonrandom treatment assignment.

We conclude this section by revisiting our SearchIt example. Rather than run an experiment, SearchIt could have simply collected data for 100,000 searches for “tablet,” recording the firm that was in the top ad position and the firm in the fourth ad position, and whether they were clicked. In this case, treatment assignment (i.e., being in top position rather than fourth position) is not random; rather, it is determined by the outcome of firms bidding in an auction. Consequently, firms who were willing to pay more for ad position are the ones that end up with the treatment.

Suppose for our nonrandom SearchIt example, it is small firms with little brand recognition who generally are willing to pay the most to be in a high ad position. How can this

nonrandom assignment of ad position cause $ETT \neq ATE$ or Selection Bias $\neq 0$? First, firms with little brand recognition may benefit the most from having a high ad position. Were they lower on the scale, their less-familiar status with customers might cause them to be overlooked. In contrast, a highly recognizable firm may garner attention (and hence a click) regardless of its ad position. If this is so, the effect of the treatment (moving from fourth to top ad position) on the treated (consisting largely of small-presence firms) is not the same as the average treatment effect (i.e., $ETT \neq ATE$). In fact, the ETT is likely more than the ATE.

Second, firms with little brand recognition likely garner fewer clicks, all else equal, than firms with high brand recognition. This means that, for a given ad position, we will see a higher likelihood of a click for a firm with high brand recognition than one with low brand recognition. Consequently, when in the fourth ad position, we expect a higher likelihood of a click for firms with high brand recognition. This means that, on average, the treated group (mostly of low-brand-recognition firms) would have lower likelihood of a click than the untreated group (consisting mostly of high-brand-recognition firms) when neither receives the treatment (i.e., they are in fourth ad position). This implies there is a non-zero selection bias. In fact, it suggests we have a negative selection bias.

To summarize the SearchIt example, the fact that low-brand-recognition firms tended to have higher ad placement resulted in both $ETT > ATE$ and $Selection\ Bias < 0$. Consequently, we again have no way of knowing whether $ETT + Selection\ Bias = ATE$ anymore; we could have $ETT + Selection\ Bias > ATE$, $ETT + Selection\ Bias < ATE$, or $ETT + Selection\ Bias = ATE$. The bottom line: Comparing the mean outcomes for the treated and untreated no longer gives us a reliable estimate of the average treatment effect, and this again is due to the nonrandom treatment assignment.

COMMUNICATING DATA 4.5

THE REWARDS OF RUDENESS

Do rude sales clerks make more sales than polite ones? It may seem counterintuitive at first that rudeness might pay off with sales. However, it is not difficult to conjure theories why this might be the case. For example, customers may respond to rudeness with big purchases to “impress” the snobby clerk. Researchers at the University of British Columbia have attempted to tackle this question in an experimental setting. Many businesses may want to learn about this question for their own products but do not have the luxury of conducting an experiment. Instead, they may gain access to customer surveys (including information on clerk rudeness) along with sales for a large number of clothing retailers. They then could compare the mean sales for stores with rude clerks against the mean sales for stores with polite clerks.

Unfortunately, comparing these two figures falls short of the experimental ideal. In particular, there is no guarantee that clerk rudeness is randomly assigned across stores. Stores with rude clerks may generally have more sales even if they had polite clerks ($\text{Selection Bias} \neq 0$) or have customers who respond to rudeness differently than those of stores with polite clerks ($\text{ETT} \neq \text{ATE}$). So, while this comparison may be interesting, it is not a reliable estimator of the true ATE from using rude clerks as opposed to polite clerks.

Interestingly, in the study that actually used an experimental setting (with randomized treatment), the researchers found that rude clerks do in fact land more big-ticket sales.

RISING TO THE dataCHALLENGE

Does Dancing Yield Dollars?

Let’s return to the Data Challenge posed at the start of the chapter: finding a way to measure the effectiveness of the dancers outside the fast-food restaurant. Understanding the ability of a properly run experiment to measure

an average treatment effect, you recommend the following course of action. You randomly pick 10 weeks out of the year. Then, across those 70 days, you randomly choose whether to have dancers in front of the store or not, and record the sales of the store for all 70 days. Lastly, you take the difference between the mean sales during the “dancing days” and the mean sales during the “nondancing days.” From [Reasoning Box 4.1](#), you know that this difference is an unbiased estimate of the actual ATE of having dancers in front of the store. You could then run a *t*-test to determine whether the ATE is significantly different from zero, and build a confidence interval to determine the plausible range of actual ATEs.

To illustrate, suppose mean sales on days with dancers was \$12,974, with standard deviation of \$1,316. Also, mean sales on days without dancers was \$12,439, with standard deviation of \$1,237. Lastly, suppose your random allocation resulted in an even split of 35 days with dancers and 35 days without dancers. You would then conclude that an unbiased estimate of the effect of the dancers is $\$12,974 - \$12,439 = \$535$. Your *t*-stat is 1.75 (> 1.65), so you are 90% confident there is an effect from the dancers. Lastly, you are 90% confident that effect is between \$31 and \$1,039.

SUMMARY

This chapter introduced the scientific method and detailed each of its components. It showed how, through experiments with random treatment assignment, we are able to reliably use the data from those experiments to measure the average treatment effect (ATE). It went on to demonstrate how to test hypotheses about the ATE and build confidence intervals for the ATE using experimental data. Lastly, it contrasted experimental and nonexperimental data, illustrating the potential consequences of using nonexperimental data in the same way as experimental data to measure an ATE.

As the title of this chapter indicates, the scientific method is the gold standard for establishing causality. Throughout the rest of the book, we will be dealing with nonexperimental data, since this is what is typically found in

business environments. Despite the data being nonexperimental, we will be attempting to establish causal relationships. To this end, the scientific method and experimental data serve as a benchmark for the analyses we will be discussing henceforth. The goal is to apply methods and reasoning that ensure our results using nonexperimental data produce estimates of causal relationships that we would have found had we been able to run an experiment.

110

KEY TERMS AND CONCEPTS

average treatment effect (ATE)

effect of the treatment on the treated (ETT)

experiment

experimental data

hypothesis

nonexperimental data

scientific method

selection bias

treatment

treatment effect

CONCEPTUAL QUESTIONS connect

1. Which of the following is *not* an element of the scientific method: (LO1)
 - a. Formulate a hypothesis
 - b. Do background research
 - c. Collect market data
 - d. Communicate the findings
2. Generate a question of causality in business, i.e., a question comparable to the one we discussed in the text: “what is the effect of increasing the size of a banner ad on the click-through rate for the advertiser?” In doing

so, you will have executed the first step in the scientific method. Then, explain how you would execute the remaining five steps of the scientific method for your question. (LO1)

3. Explain the difference between the average treatment effect (ATE) and the effect of the treatment on the treated (ETT). (LO2)
4. A local store manager is pondering implementing a 10% across-the-board price increase for her store the following day but wonders what the effect on her profits will be. To answer this, she uses data from two prior days, where one had prices as they are now and the other had the 10% price increase. She notes that profits were \$1,532 on the day with current prices, and \$1,787 on the day with the 10% higher prices. She then determines that increasing price tomorrow by 10% will raise profits by \$255 ($1,787 - 1,532$). (LO2)
 - a. Explain why her method of measuring the effect of the price increase is flawed.
 - b. If she wanted an accurate measurement of the effect of the price increase on tomorrow's profits, what information would she need? (Hint: It's not physically possible to get it.)
5. Which of the following is an example of selection bias equaling zero? (LO2)
 - a. Older men respond more strongly to a new drug than younger men.
 - b. When considering making a price change on Wednesday, a manager notes that at the current price of \$10, sales on Monday are, on average, the same as sales on Wednesday.
 - c. Customers who receive a coupon buy your product at the same rate as customers who do not.
6. Which of the following is an example where $ETT \neq ATE$? (LO2)
 - a. Customers who receive a coupon are more price-sensitive than customers who do not.
 - b. Patients who received a drug were more likely to get sick than those who didn't, if neither group were to receive the drug.
 - c. At the current price of \$10, sales on Monday are, on average, the same as sales on Wednesday.

7. What is the primary reason that using nonexperimental data to measure a treatment effect can be problematic? (LO5)
8. Concisely explain why business data typically involves nonrandom assignment of strategic variables. (LO5)

111

-
9. Suppose you have data on 200 firms, and half advertise on Google. If the advertising firms had higher sales than nonadvertising firms before they started advertising on Google, does this fact impact your ability to measure the effect of Google advertising on sales using these data? (LO6)
 10. Should we expect firms to selectively advertise in a way that makes $ETT > ATE$, $ETT = ATE$, or $ETT < ATE$? (LO6)

QUANTITATIVE PROBLEMS connect

11. To date, you have staunchly avoided placing any advertisements on your website. However, with revenues declining, you are considering relenting on this position. Before making a full commitment, you decide to try to determine whether the presence of ads has any negative impacts on your sales. Therefore, you make a deal with an advertiser to show a 10-second pop-up ad (that pops up and plays when the site is first visited and then becomes a banner ad on the page) intermittently over a period of one month. Suppose you've decided to show the ad during business hours (9:00 A.M.– 5:00 P.M.), and not show the ad any other time during the month. (LO2)
 - a. What are the “subjects” in this (field) experiment?
 - b. What is the treatment?
 - c. What is the relevant outcome?
 - d. Is treatment assignment random?
 - e. Given the way treatment is assigned, is there reason to believe:
 - i. $ETT \neq ATE$?
 - ii. Selection Bias $\neq 0$?
12. Refer to Problem 11, and suppose for each visitor to your site, whether the visitor sees the ad is determined randomly. After the month is completed,

you have the following information:

Mean sales when ad was shown: \$27.67

Standard deviation of sales when ad was shown: \$19.13

Number of times ad was shown: 8,172

Mean sales when ad was not shown: \$28.21

Standard deviation of sales when ad was not shown: \$19.82

Number of time ad was not shown: 10,437

Does showing the ad affect your sales? Explain your reasoning. (LO3)

13. Using the data from Problem 12, the advertiser claims there is evidence that running the ads may actually improve your sales. Is there evidence for this? Explain your reasoning, why or why not. (LO4)
14. Suppose you have run an experiment and afterward came to realize that your treatment assignment was not random. You had 200 participants, 100 receiving the treatment and 100 not receiving the treatment. You have the following figures from the experiment: (LO2)

$$\overline{\text{Outcome} \mid \text{Treatment} = 1} = 7.2$$

$$\overline{\text{Outcome} \mid \text{Treatment} = 0} = 6.3$$

If you are confident there is no selection bias, what can these figures estimate for you? Provide the estimate.

112

-
15. Jim owns his own burger restaurant. On some days he posts a sign that advertises his special Jimbuger, and on other days he does not. Jim has been collecting data on Jimburger sales for the past 70 days, with the following results: (LO3)
 - For 38 of those days, Jim posted his Jimburger advertisement; the mean unit sales were 89 with a standard deviation of 17.
 - For 32 of those days, Jim did not post the advertisement; the mean unit sales were 77 with a standard deviation of 15.
 - a. If we treat the 70 days Jim collected data as a random sample, and if Jim randomly chose when to advertise, what is the distribution of the difference between mean unit sales when advertising and

mean unit sales when not advertising?

- b. Suppose Jim believed that the advertisement was responsible for an increase in sales of Jimburgers by 20 burgers. Do these data support or reject this hypothesis?
16. A regional Internet service provider (ISP) is interested in whether a 15% discount on its price for its highest-speed service will notably impact customer retention. To answer this question, the ISP randomly chooses 75 of the 150 markets it serves to receive the 15% discount, and keeps prices the same in the remaining 75 markets. After six months, the ISP observes the following: (LO4)
- For the markets receiving the discount, the average retention rate was 0.93 (i.e., 93%), with a standard deviation of 0.255.
 - For the markets not receiving the discount, the average retention rate was 0.88, with a standard deviation of 0.325.

Using a 95% confidence interval, determine whether you can state with 95% confidence that the discount improved the six-month retention rate.

Linear Regression as a Fundamental Descriptive Tool

5

LEARNING OBJECTIVES

After completing this chapter, you will be able to:

- LO5.1** Construct a regression line for a dichotomous treatment.
- LO5.2** Construct a regression line for a multi-level treatment.
- LO5.3** Explain both intuitively and formally the formulas generating a regression line for a single treatment.
- LO5.4** Distinguish the use of sample moment equations from estimation via least squares.
- LO5.5** Distinguish regression equations for single and multiple treatments.
- LO5.6** Describe a dataset with multiple treatments using multiple regression.
- LO5.7** Explain the difference between linear regression and a regression line.

Chapter opener image credit: ©naqiewei/Getty Images

dataCHALLENGE Where to Park Your Truck?

You just purchased a food truck, and have begun selling in a large college town. As you are learning the market, you have been changing location every few days to get a sense of local demand. In doing so, you've decided to collect some data. In particular, you collect data on your revenues and the distance of the truck location from the center of the local university. The hope is to get a sense as to whether demand notably varies depending on your proximity to the university. The data you collected thus far are shown in [Table 5.1](#).

114

TABLE 5.1 Food Truck Data on Revenue and Distance to University

DATE	REVENUE (\$)	DISTANCE (MILES)
9/13	750	1.2
9/14	835	1.2
9/15	694	2.4
9/16	558	2.4
9/17	732	2.4
9/20	906	3.3
9/21	632	3.3
9/22	817	0.4
9/23	916	0.4
9/24	688	0.4
9/27	801	0.8
9/28	582	0.8
9/29	733	1.7

9/30	940	1.4
10/1	608	2.7
10/4	816	0.5
10/5	775	0.5
10/6	590	2.0
10/7	765	2.0
10/8	782	1.1

Using these data, describe the relationship between revenue and distance to the university.

Introduction

In [Chapter 4](#), we described the scientific method primarily in terms of a dichotomous treatment, in that experiment participants either receive the treatment or they do not. However, treatments can take different forms (dichotomous and multi-level), and it is also possible to experience more than one treatment at a time. While our ultimate goal is to establish causal effects of treatments, to achieve that goal we must first establish a general way of describing the relationship between an outcome and treatment(s) of any kind. In this chapter, we introduce linear regression as a ubiquitous analytical tool for accomplishing this task. Then, in [Chapter 6](#), we distinguish when linear regression serves only as a descriptor and when it is informative about causal effects.

115

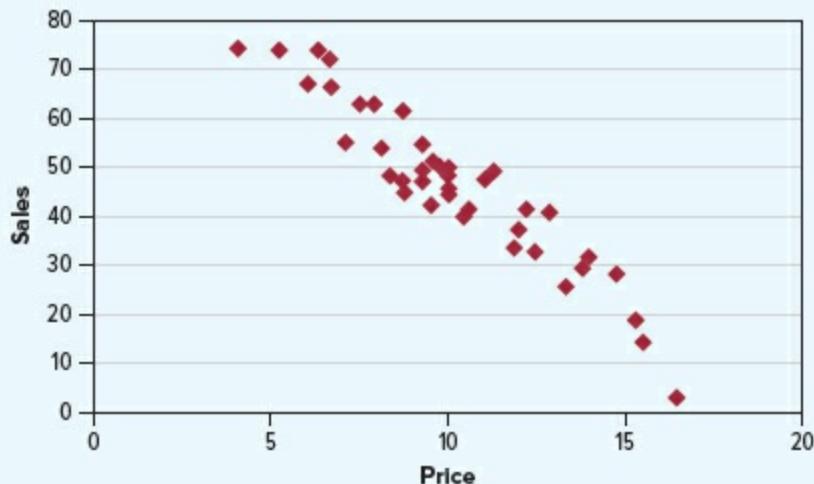
When describing the linear regression model, most books jump straight to what's known as ordinary least squares (OLS) as the estimation method, and also defer discussion of dichotomous treatments/variables until much later in

the text. We deviate from those approaches here, and for important reasons rooted in our goal of establishing causal relationships. By starting with a dichotomous treatment, we are able to build the regression line using exactly the same measurements we used for the scientific method; thus, we establish a natural link to our foundational discussion in [Chapter 4](#) and a highly intuitive basis for the regression line's construction. In addition, we introduce *moment conditions*, rather than OLS, as the foundation for estimating linear regression models. We use moment conditions because they greatly simplify the reasoning process for establishing causality, which we discuss in detail in [Chapter 6](#). For those who have seen regression before with a focus on OLS, we note that the material here is not at odds with what you've seen. We show that using moment conditions leads to the same solutions as OLS, but with much stronger conceptual ties to the scientific method and the process of establishing causality.

To get a sense of what we mean by describing a relationship, consider the following example. Suppose we have data on price and sales for a given firm. Here, sales is the outcome and price takes the form of a multi-level treatment, detailed further later in the chapter. For now, suppose [Figure 5.1](#) is a scatterplot of our price and sales data.

With this graph in hand, how do we summarize the relationship between these two variables? As we move along the X-axis (as price increases), we tend to move down the Y-axis (sales fall). So, we may summarize the relationship as a negative one. But can we say more? On average, how much do sales appear to fall when price is one dollar

FIGURE 5.1 Scatterplot of Price and Sales



116

higher? Or, in other words, what is the average rate of change in the outcome with a change in the treatment?

Answering such a question really just requires us to know something about the slope of the relationship between Y and X . The simplest way to get this measure is to draw, and solve for, a line that we believe best describes the data we observe. In this chapter, we will detail how to solve for lines (and other functions) that best describe the data for two-dimensional cases as in [Figure 5.1](#), and for multi-dimensional cases as well.

It is important to note that our objective in this chapter is to *describe* the relationship between a treatment(s) and an outcome. While accomplishing this objective can be informative and even enlightening, it is not equivalent to measuring treatment effects or characterizing causal relationships between variables. To take that additional step, we will need to integrate some assumptions and reasoning, which we will do in the next chapter.

The Regression Line for a Dichotomous Treatment

AN INTUITIVE APPROACH

LO 5.1 Construct a regression line for a dichotomous treatment.

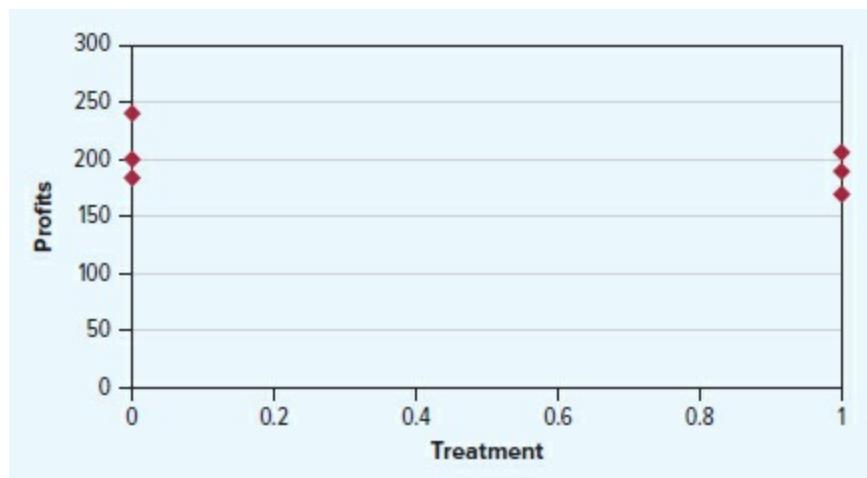
To begin, let's consider a simplified business scenario. Suppose Jill sells corn at the local farmers' market each Saturday. Each week, she considers exactly two prices for an ear of corn at her booth: \$1.00 and \$1.50. At the end of each Saturday, she records the price she charged and the profits she made for that day. After six weeks, her data look as in [Table 5.2](#).

For our example with Jill, there are only two treatment statuses: a price of \$1.00 and a price of \$1.50. Consequently, we can consider the weeks where price was \$1.00 as being untreated and weeks where price was \$1.50 as being treated. Characterizing the pricing difference this way implies that the “treatment” is a price increase of \$0.50. (Of course, we could have taken the reverse approach, where weeks when price was \$1.50 are untreated and weeks when price was \$1.00 are treated; this would simply give a mirror image of our results and not qualitatively affect our findings.)

TABLE 5.2 Price and Profits for Jill's Corn

PRICE	PROFITS
\$1.00	\$240
\$1.00	\$200
\$1.00	\$185
\$1.50	\$205
\$1.50	\$170
\$1.50	\$190

FIGURE 5.2 Scatterplot of Profits and Treatment Status



Since there are just two treatment statuses—treated and untreated—this is an example of a **dichotomous treatment**. Our discussion of experiments in [Chapter 4](#) focused on these types of treatments, e.g., participants either receive a drug that treats cancer or they do not. We use this type of treatment as our starting point for building a general method of describing the relationship between an outcome and a treatment(s). We will consider alternative treatment types later in this chapter.

dichotomous treatment Two treatment statuses—treated and untreated.

Letting weeks where price is \$1.00 be untreated (Treatment = 0, meaning the treatment of a \$0.50 price increase was not given) and weeks where price is \$1.50 be treated (Treatment = 1, meaning the treatment of a \$0.50 price increase was given), we can plot the data from [Table 5.2](#), as presented in [Figure 5.2](#).

Suppose now that we want to draw a line through these data that we believe best describes the relationship between Profits and Treatment implied by the data points in [Figure 5.2](#). To do so is to engage in what's known as **regression analysis**, the process of using a function to describe the relationship among variables. For our simple example, the function we seek is a line that describes the relationship between Profits and Treatment.

regression analysis The process of using a function to

describe the relationship among variables.

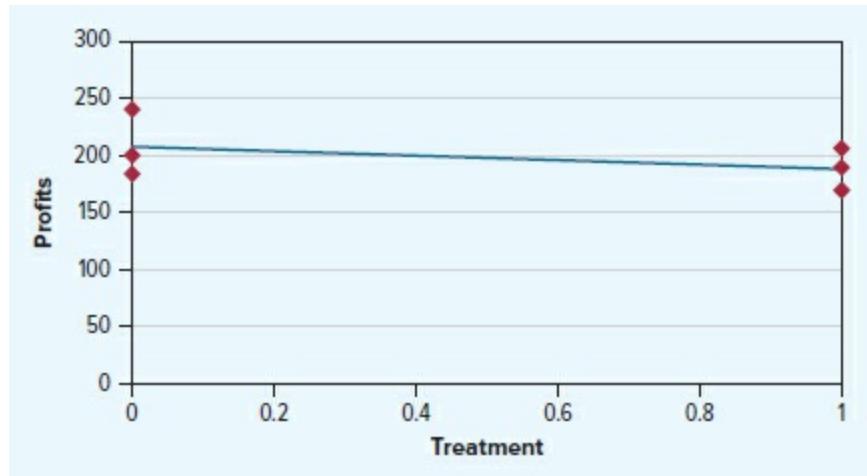
Is there an intuitive way to construct this line? In general, the formula for a line is:

$$Y = f(X) = b + mX$$

Here, b is the intercept and m is the slope of the line. In this simplified scenario, constructing our “best” line essentially comes down to plotting two points in the graph— $f(0)$ and $f(1)$ —and then connecting those two points. This is because there are data points for only two values on the X -axis: 0 and 1.

Let’s start by choosing a value for $f(0)$. For our example, this means we must choose a value of Profits that best describes the data points we observed when the treatment was not given. Perhaps the most intuitive choice is the mean of profits when the Treatment was not given. Here, the mean of profits for the untreated weeks was \$208.33, and so we set $f(0) = 208.33$. Applying the same reasoning for $f(1)$ leads us to choose the mean of profits when the Treatment was given; hence, $f(1) = \$188.33$. We then plot these two points and connect them. This gives us a line that describes the relationship between Profits and Price in our data. We show this line in [Figure 5.3](#).

FIGURE 5.3 Line Describing the Relationship Between Profits and Treatment



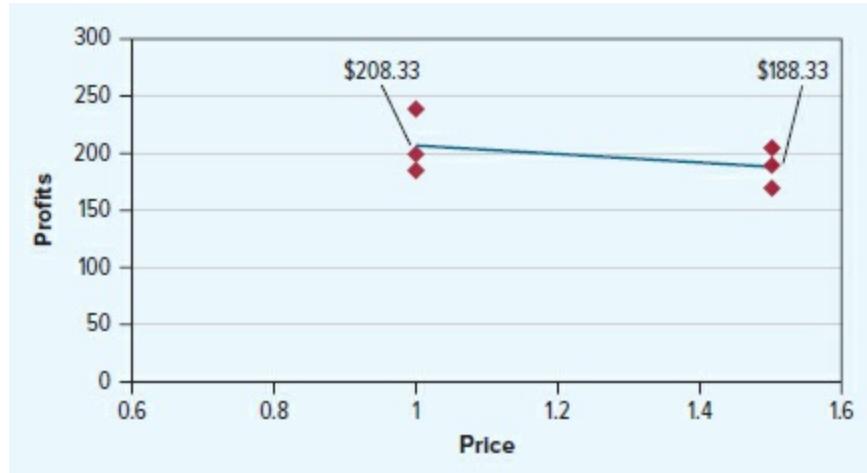
What is the equation for the line we have drawn in [Figure 5.3](#)? We know $f(0)$ and $f(1)$, but what are m and b ? First, b is simply the Y -intercept, which is defined as $f(0)$. Hence, $b = 208.33$. We can find the slope, m , by taking any two points on the line and dividing the difference in their Y values by the difference in their X values (that is, take “rise over run”). Here, we have just two points, $f(0)$ and $f(1)$, so we can take $188.33 - 208.33$ (difference in Y values) and divide by $1 - 0$ (difference in X values). Thus, $m = \frac{188.33 - 208.33}{1 - 0} = -20$. Putting everything together, the equation of the line that describes our corn profit data is:

$$\text{Profits} = 208.33 - 20 \times \text{Treatment}$$

While linking the outcome to the treatment in the form of a line is the most basic approach we can take, it is often more practical to restate the line in terms of the treatment’s units. In our example, we can restate the line in terms of price levels. In doing so, we can again graph our data, but replace Treatment with Price on the X -axis, as in [Figure 5.4](#). With this formulation, we again have two points, but they are $f(1.00)$ and $f(1.50)$, where $f(1.00) = 208.33$ and $f(1.50) = 188.33$. We connect these points to get our line relating Profits to Price, as shown in [Figure 5.4](#).

FIGURE 5.4 Line Describing Relationship Between Profits and

Price



119

Knowing these two points on our Profits/Price line, we solve for the slope and intercept. The slope again is rise over run, i.e., $m = \frac{188.33 - 208.33}{1.50 - 1.00} = -40$. The intercept, b , is $f(0)$. For this formulation, we don't have $f(0)$ directly, as we did when Treatment was on the X -axis; however, solving for $f(0)$ involves moving one unit (dollar) backward on our line from $f(1.00)$. Therefore, $f(0) = f(1.00) + (-1) \times m = 208.33 + 40 = 248.33$. Putting this all together, our line relating Profits to Price is:

$$\text{Profits} = 248.33 - 40 \times \text{Price}$$

To conclude this subsection, we can generalize the approach we took with our Profits/Price corn example. Whenever there is a dichotomous treatment—meaning that every observation either received the treatment (Treatment = 1) or did not (Treatment = 0)—we can build a line describing the relationship between the treatment and outcome by using the means for each treatment status. In particular, we calculate the mean outcome for the treated group ($\overline{\text{Outcome} | \text{Treated} = 1}$) and the mean outcome for the

untreated group $(\overline{\text{Outcome} \mid \text{Treated} = 0})$. Then, to construct our line, we start by setting $f(0) = (\overline{\text{Outcome} \mid \text{Treated} = 0})$ and $f(1) = (\overline{\text{Outcome} \mid \text{Treated} = 1})$. The equation for the line is:

$$\text{Outcome} = \overline{\text{Outcome} \mid \text{Treated} = 0} + (\overline{\text{Outcome} \mid \text{Treated} = 1} - \overline{\text{Outcome} \mid \text{Treated} = 0}) \cdot \text{Treated}$$

The above equation is the **regression line for a dichotomous treatment**. This is a special case, and the simplest case, of the simple regression line, detailed below.

regression line for a dichotomous treatment For a dichotomous treatment, the line describing the relationship between the treatment and outcome by using the means for each treatment status.

Note how the regression line for a dichotomous treatment naturally links to our formula for measuring the treatment effect from [Chapter 4](#). The slope of this line is the difference in mean outcomes between the treated and untreated. As we know from [Reasoning Box 4.1](#), this difference is an unbiased estimate of the treatment effect when participants (in this case, weeks) are a random sample and the treatment is randomly assigned. Hence, if we are willing to make these assumptions, the slope of the regression line provides an unbiased estimate of the treatment effect. Without these assumptions, the regression line serves only as a descriptive tool, describing how the outcome and treatment status move together in the data.

A FORMAL APPROACH

Rather than relying just on intuition (i.e., use the natural choice of the mean outcome to plot the points for each treatment status), we can follow a more formal approach toward constructing the regression line for a dichotomous treatment. Adding some rigor to the process will facilitate our ability to

conduct, and reason with, regression analysis for more complex treatments, where it is more difficult for us to directly build on intuition per se.

As before, we can construct the line for our Profits/Price example by choosing $f(1.00)$ and $f(1.50)$ —the points on the line corresponding to being untreated and treated, respectively—and then solving for the slope (m) and intercept (b). Before we make our choices for $f(1.00)$ and $f(1.50)$, let's define our observed outcomes in terms of these two points on the line.

$$\text{Profit}_i = f(1.00) + e_i \text{ if Price}_i = 1.00$$

$$\text{Profit}_i = f(1.50) + e_i \text{ if Price}_i = 1.50$$

120

5.1 Demonstration Problem

Suppose you are interested in the relationship between a person's annual salary and whether they have a college degree. You have collected data for 16 individuals, shown in [Table 5.3](#). Letting the acquisition of a college degree be the Treatment, solve for the "regression line for a dichotomous treatment" for these data.

TABLE 5.3 Salary and Indicator of College Degree for 16 Individuals

INDIVIDUAL NUMBER	SALARY	COLLEGE DEGREE
1	\$28,000	No
2	\$42,000	No
3	\$59,000	Yes
4	\$37,000	No
5	\$81,000	Yes
6	\$106,000	Yes

7	\$72,000	No
8	\$23,000	No
9	\$41,000	No
10	\$38,000	Yes
11	\$35,000	No
12	\$62,000	Yes
13	\$49,000	No
14	\$30,000	No
15	\$56,000	Yes
16	\$27,000	No

Answer:

For this problem, salary is the Outcome and the acquisition of a college degree is the Treatment. To solve for the regression line, we must calculate the average salary for individuals without a college degree ($\overline{\text{Outcome} \mid \text{Treated} = 0}$) and the average salary for individuals with a college degree ($\overline{\text{Outcome} \mid \text{Treated} = 1}$). These calculations are \$38,400 and \$67,000, respectively. Therefore, the equation of the corresponding regression line is:

$$\text{Salary} = 38,400 + \$28,600 \times \text{Degree}$$

where Degree equals one if the individual acquired a college degree and zero otherwise.

Here, the subscript i simply delineates different observations. So, for our pricing example, i takes on the values one through six ($i \in \{1, 2, \dots, 6\}$), since there are six observations.

TABLE 5.4 Residuals for Price of \$1.00 when $f(1.00) = \$220$

PRICE	PROFITS	$f(1.00)$	RESIDUAL
\$1.00	\$240	\$220	20
\$1.00	\$200	\$220	-20
\$1.00	\$185	\$220	-35

In the above formulations, e_i is the residual for observation i . The **residual** is defined as the difference between the observed outcome and the corresponding point on the regression line for a given observation. In general terms, for a given X_i , the residual is:

$$e_i = Y_i - f(X_i)$$

Within our Profit/Price example then, for a given Price $_i$, $e_i = \text{Profit}_i - f(\text{Price}_i)$.

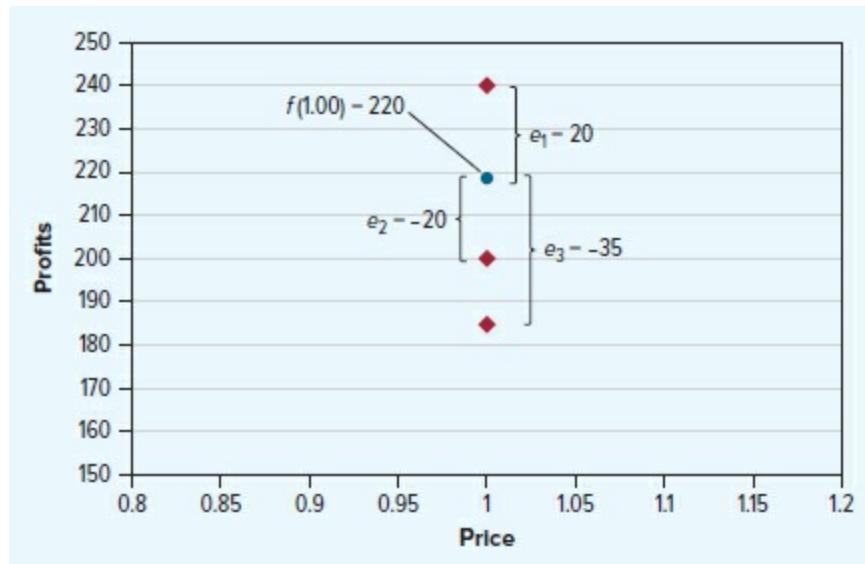
residual The difference between the observed outcome and the corresponding point on the regression line for a given observation.

Using the above framework with residuals, let's choose values for $f(1.00)$ and $f(1.50)$ that we believe best describe the data. Consider an arbitrary guess for $f(1.00)$. Suppose we choose $f(1.00) = 220$. That is, for a price of \$1.00, our line indicates that profit of \$220 best describes the data for that price point. This choice implies three residuals, since we have three observations when price is \$1.00. From [Table 5.2](#), we observe: (1.00, 200), (1.00, 240), and (1.00, 185). The three residuals are as calculated in [Table 5.4](#) and illustrated in [Figure 5.5](#).

Looking at these residuals, does it appear that $f(1.00) = \$220$ best describes the data? To answer that question, consider each residual for the observations when price is \$1.00:

- As we see in [Figure 5.5](#), the first residual is 20. This means that the actual profits we observed (240) were 20 higher than the point on our line (220) for this observation. Thus, the point on our line “undershoots” the actual profits in this case.
- The second residual is -20 . This means the actual profits we observed (200) were 20 lower than the point on our line (220) for this second observation. Thus, the point on our line “overshoots” the actual profits in this case.
- Lastly, our third residual is -35 . This means the actual profits we observed (185) were 35 lower than the point on our line (220) for this final observation. Thus, the point on our line again overshoots the actual profits.

FIGURE 5.5 Scatterplot of Residuals for Price of \$1.00 when $f(1.00) = \$220$



122

We can see from [Table 5.4](#) that if $f(1.00) = 220$, then the average residual is $[20 + (-20) + (-35)]/3 = -11.67$. This implies that, on average, the point we've chosen for the line tends to overshoot the data. For this reason, we may conclude that it is not our best choice to describe the data when price is

\$1.00.

But what choice for $f(1.00)$ would best describe the data? Intuitively, a choice for $f(1.00)$ is best if it tends to neither overshoot nor undershoot the observed outcomes. More formally, a choice for $f(1.00)$ is best if its corresponding residuals are, on average, zero.

Using this definition of “best” for a point on our line, let’s solve for the best choice for $f(1.00)$ in our example. For the residuals to average zero, this means:

$$\frac{(\text{Profits}_1 - f(1.00)) + (\text{Profits}_2 - f(1.00)) + (\text{Profits}_3 - f(1.00))}{3} = 0$$

Plugging in for the observed values of Profits and solving for $f(1.00)$, we have:

$$f(1.00) = \frac{200+240+185}{3} = 208.33$$

Hence, our best choice for $f(1.00)$ is the average of Profits when Price is \$1.00, which equals 208.33. Similarly, our best choice for $f(1.50)$ is the average of Profits when Price is \$1.50, which equals $(205 + 170 + 190)/3 = 188.33$.

In general terms, for a given outcome and a dichotomous treatment, we can define “best” as having residuals that average zero both for the treated and untreated observations. Then, the best choices for the points on our line corresponding to being untreated and treated are their corresponding average outcomes, i.e., $\left(\overline{\text{Outcome} | \text{Treated} = 0}\right)$ and $\left(\overline{\text{Outcome} | \text{Treated} = 1}\right)$, respectively.

Notice that these choices are exactly the ones we made following only an intuitive approach in the prior subsection. Consequently, they again lead us to the same regression line for a dichotomous treatment:

$$\text{Outcome} = \overline{\text{Outcome} | \text{Treated} = 0} + \left(\overline{\text{Outcome} | \text{Treated} = 1} - \overline{\text{Ou}} \right)$$

We summarize the basic reasoning of this section in [Reasoning Box 5.1](#).

REASONING BOX 5.1

THE REGRESSION LINE FOR A DICHOTOMOUS TREATMENT

For the case of a dichotomous treatment, define a line as best describing the data if it generates residuals that average zero for both the untreated and treated observations. Then, this line will contain the points:

$(0, \overline{\text{Outcome} | \text{Treated} = 0})$ and $(1, \overline{\text{Outcome} | \text{Treated} = 1})$. The full equation for the line best describing the data is:

$$\text{Outcome} = \overline{\text{Outcome} | \text{Treated} = 0} + (\overline{\text{Outcome} | \text{Treated} = 1} - \overline{\text{Outcome} | \text{Treated} = 0}) \cdot \text{Treated}$$

This is defined as the regression line for a dichotomous treatment.

123

5.2

Demonstration Problem

Consider the following dataset in [Table 5.5](#) containing information on a Treatment and Outcome. Next, consider two possible lines describing these data. Line 1 passes through the points $(0, 20)$ and $(1, 32)$, and Line 2 passes through the points $(0, 36)$ and $(1, 40)$. Explain both formally and intuitively why neither Line 1 nor Line 2 best describes these data. Then, solve for the line that does best describe these data.

TABLE 5.5 Treatment and Outcome Data

TREATMENT	0	1	1	0	1	0	0	1	0	1
OUTCOME	20	30	50	10	60	20	40	70	50	30

Answer:

For Line 1, the residuals for the untreated observations have an average value of:

$$\frac{0-10+0+20+30}{5} = 8$$

The residuals for the treated observations have an average value of:

$$\frac{-2+18+28+38-2}{5} = 16$$

Intuitively, the proposed points for the untreated and the treated both undershoot the data. The residuals for neither choice average zero, ensuring these choices are not best.

For Line 2, the residuals for the untreated observations have an average value of:

$$\frac{-16-26-16+4+14}{5} = -8$$

The residuals for the treated observations have an average value of:

$$\frac{-10+10+20+30}{5} = 8$$

Intuitively, the proposed point for the untreated overshoots the data, and the proposed point for the treated undershoots the data. Again the residuals for neither choice average zero, ensuring these choices are not best.

The best line generates residuals that average zero for both the treated and untreated observations. The average outcome for the untreated is 28, and the average outcome for the treated is 48. Therefore, the best line passes through the points (0, 28) and (1, 48). The equation for this line is:

$$\text{Outcome} = 28 + 20 \times \text{Treatment}$$

124

The Regression Line for a Multi-Level Treatment

LO 5.2 Construct a regression line for a multi-level treatment.

LO 5.3 Explain both intuitively and formally the formulas generating a regression line for a single treatment.

AN INTUITIVE APPROACH

Many treatments in business, medicine, and beyond come in more than just one level. For example, in medicine, we may consider not only the effect on health outcomes from taking a drug versus not; we may also consider the effect on health outcomes from taking different *dosage levels* of the drug. Analogously, in business, we generally consider not just the effect on profits from charging a single high price versus a single low price; instead, we are interested in the effect on profits from charging various different prices. When a treatment can be administered in more than one quantity, we say it is a **multi-level treatment**.

multi-level treatment A treatment that can be administered in more than one quantity.

Let's again consider our example of Jill selling corn at the local farmers' market, but this time let's extend her pricing possibilities. Now, Jill can charge one of three prices in a given week: \$1.00, \$1.50, or \$2.00. This

simple addition of \$2.00 as another price point means price is no longer a dichotomous treatment; instead, it is a multi-level treatment. As in the dichotomous example, we could characterize the price of \$1.00 as being untreated and a \$0.50 price increase as the treatment. Then, charging a price of \$1.50 is the equivalent of administering one “dose” of the treatment, and charging a price of \$2.00 is the equivalent of administering two “doses” of the treatment. Thus, we could create a Treatment variable that is 0 when price is \$1.00, 1 when price is \$1.50, and 2 when price is \$2.00. In practice, though, the treatment variable is seldom explicitly converted into treatment doses for multi-level treatments; rather, the unit of measurement for the treatment variable implicitly is the dosage. In our Price/Profits example, dosage is measured in dollars.

Consider now a new dataset for Jill with nine weeks of data, presented in [Table 5.6](#). We plot these data in [Figure 5.6](#).

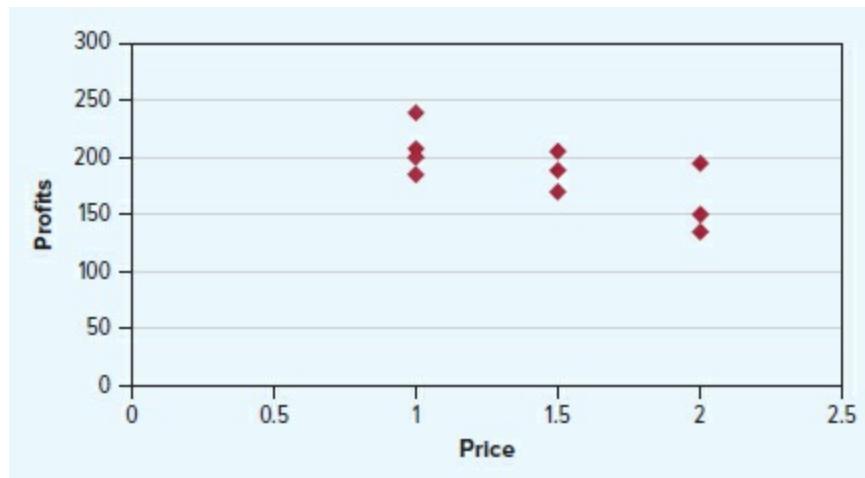
Suppose we again want to draw a line that we believe best describes the relationship between Profits and Price implied by the data points in [Figure 5.6](#). Can we follow the approach we used for a dichotomous treatment? If we follow that approach, we would plot $f(1.00)$, $f(1.50)$, and $f(2.00)$. And if we force the residuals to average zero for each price, we would choose the average profits when price was \$1.00 for $f(1.00)$, the average profits when price was \$1.50 for $f(1.50)$, and the average profits when price was \$2.00 for

TABLE 5.6 Price and Profits for Jill’s Corn with Three Price Levels

PRICE	PROFITS
\$1.00	\$240
\$1.00	\$200
\$1.00	\$185
\$1.50	\$205
\$1.50	\$170
\$1.50	\$190

\$2.00	\$195
\$2.00	\$150
\$2.00	\$135

FIGURE 5.6 Scatterplot of Profits and Price for Jill's Corn with Three Price Levels



$f(2.00)$. We then connect these three points to form a line, and solve for the slope (m) and intercept (b) of that line.

Unfortunately, the approach we used for the dichotomous treatment generally does not work for a multi-level treatment. The problem is that, when we plot three or more points on a graph, it is generally the case that they will not fall on the same line. For our Price/Profits example, if we plot our three points using average profits, we have $f(1.00) = 208.33$, $f(1.50) = 188.33$, and $f(2.00) = 160$. We plot these three points (in red) in [Figure 5.7](#), in addition to the nine data points on Price and Profits.

Notice that it is not possible to connect all three points on a single line; we provide one attempt in [Figure 5.7](#). In fact, using the average outcome to plot the points for each treatment level generally will result in this problem when there are more than two treatment levels. The reason lies in the fundamentals of linear algebra. If we have the coordinates for points on a

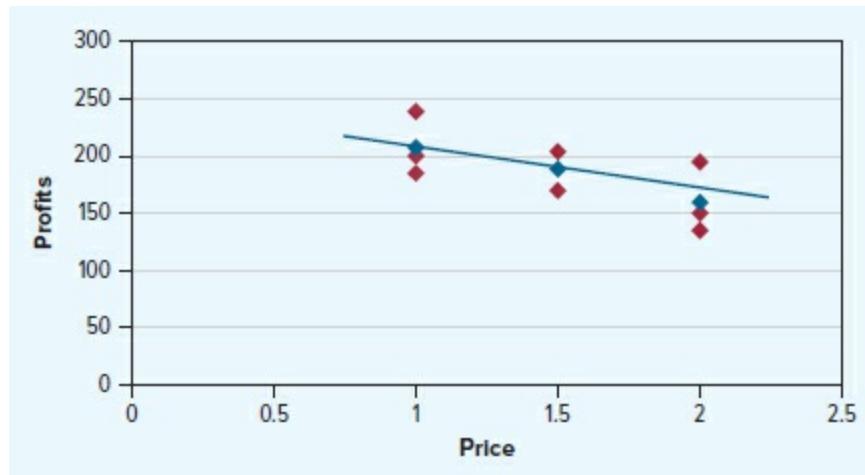
line, we can solve for the slope and intercept by plugging in the coordinates to the general linear equation. In our example, we have three points, resulting in three equations:

$$f(1.00) = b + m \times 1.00$$

$$f(1.50) = b + m \times 1.50$$

$$f(2.00) = b + m \times 2.00$$

FIGURE 5.7 Line Attempting to Connect Average Profits for Three Price Levels



126

Plugging in the average profits for each price, we have:

$$208.33 = b + m \times 1.00$$

$$188.33 = b + m \times 1.50$$

$$160 = b + m \times 2.00$$

Unfortunately, we cannot solve for m and b . This is because we have three equations to solve but only two “unknowns” with which to do it. For example, we know $m = -40$ and $b = 248.33$ solves the first two equations; however, when we plug these values into the third, the equality does not hold. In general, except in very rare cases, we will not be able to find a solution when there are more equations than unknowns. And this will be the case whenever we have a multi-level treatment and attempt to build the line in the same way we did for a dichotomous treatment.

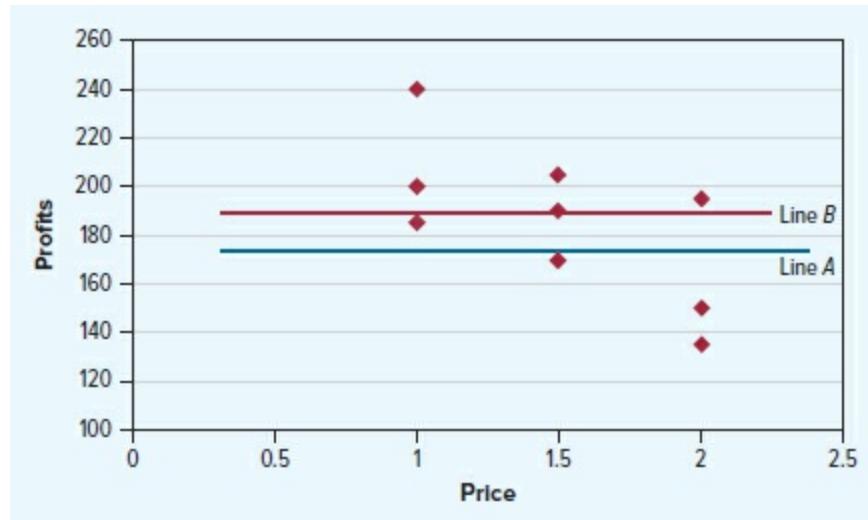
The upshot of the above discussion is that, for a multi-level treatment, we can no longer hope to build a line describing the data that has residuals averaging zero for each treatment level. Consequently, we must consider a different approach. Rather than plot an “ideal” point for each treatment level and then solve for the corresponding slope and intercept, we can try to directly solve for the slope and intercept of the line we believe best describes the data. To do this, we must have an understanding of what makes a line “best.” While the generally accepted notion of how to determine the “best” line to describe the data has formal underpinnings (detailed in the next subsection), it is also highly intuitive. The intuition is perhaps best conveyed visually.

In [Figure 5.8](#), we again plot the data from our Price/Profits example with three price points. In addition, we include two candidate lines to describe these data: Line A and Line B. If you were forced to choose from just these two lines to best describe these data, which line would you choose?

Consider the following intuitive case explaining why Line B is “better” than Line A: Notice that, for Line A, most of the data points lie above the line; only a few lie below it. In contrast, for Line B, the data points are reasonably balanced above and below the line. In short, Line A tends to undershoot the data on average. Put another way, if we randomly

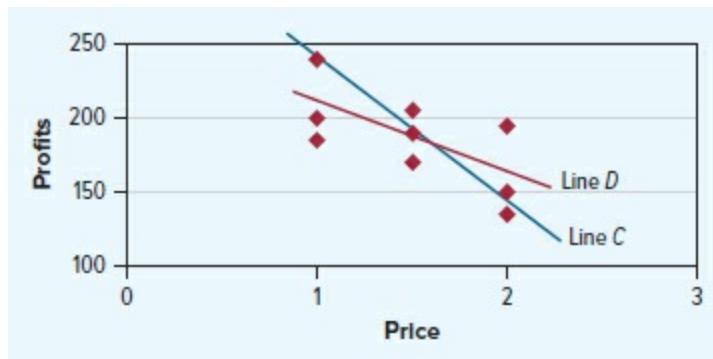
FIGURE 5.8 Two Candidate Lines for Describing Profits and

Price Data



127

FIGURE 5.9 Two More Candidate Lines for Describing Profits and Price Data



selected an observed Price/Profit combination in the data, it is likely the observed profit is higher than the corresponding point on Line A for that Price. This is not the case for Line B. It neither tends to overshoot nor undershoot the data. In sum, one way of ranking candidate lines is whether they have a tendency to generally overshoot or undershoot the data, preferring one that does neither.

In Figure 5.9, we again plot the data from our Price/Profits example with

three price levels, and present two new candidate lines to describe these data: Line C and Line D. If you were forced to choose from just these two lines to best describe these data, which line would you choose?

Consider the following intuitive case explaining why Line D is “better” than Line C: First, notice that, unlike the case for Line A in [Figure 5.8](#), neither Line C nor Line D tends to overshoot or undershoot the data. For both lines, the data are reasonably balanced above and below. However, we can make the case that Line C has an overstated slope. That is, it suggests that profits fall with price at a greater rate than the data points themselves do.

What is the source of this overstated slope? Notice that for the lowest price (\$1.00), Line C tends to overshoot the data, and at the same time, for the highest price (\$2.00), Line C tends to undershoot the data. If profits are overstated when price is low and understated when price is high, this will exaggerate the rate at which profits fall with price relative to what we saw in the data; hence, the slope is overstated (it is too steep). In contrast, for Line D, there is no obvious relationship between that line’s tendency to over- or undershoot the data and the price level. This precludes it from exaggerating or understating the rate at which profits fall with price relative to what we saw in the data.

In sum, another way of ranking candidate lines is whether their tendency to over- or undershoot the data depends on the level of the treatment (e.g., price), preferring one where this is not the case.

Thus, we have just constructed two intuitive criteria for a line to best describe the data: (1) It should not generally overshoot or undershoot the data, and (2) its tendency to over- or undershoot the data across specific price levels should not depend on the price level. Next, we formalize these criteria.

A FORMAL APPROACH

We can build on this intuition to formally establish what makes a line best describe the data, and then solve for its slope and intercept. As we did in the previous section, let’s define our observed outcomes in terms of their corresponding points on the line and residuals.

For our multi-level Price/Profit example, we have three price points and nine observations. If we express points on the line in terms of the slope and intercept, this gives us:

$$\text{Profit}_i = b + m \times 1.00 + e_i \text{ if Price}_i = 1.00$$

$$\text{Profit}_i = b + m \times 1.50 + e_i \text{ if Price}_i = 1.50$$

$$\text{Profit}_i = b + m \times 2.00 + e_i \text{ if Price}_i = 2.00$$

Here, i takes on the values one through nine ($i \in \{1, 2, \dots, 9\}$), since there are nine observations. Note again that the residuals are the difference between the observed Profit and the corresponding point on the line for a given observation. Using our above formulation, then, we can write the residual for a given observation as follows:

$$e_i = \text{Profit}_i - b - m \times \text{Price}_i$$

If we followed the approach we used for a dichotomous treatment, then we would solve for the “best” line by finding a slope and intercept that makes the residuals average zero for each price point. Formally, we would find m and b such that:

$$\begin{aligned}\frac{\sum_{i=1}^3 e_i}{3} &= \frac{\sum_{i=1}^3 (\text{Profit}_i - b - m \times 1.00)}{3} = 0 \\ \frac{\sum_{i=4}^6 e_i}{3} &= \frac{\sum_{i=4}^6 (\text{Profit}_i - b - m \times 1.50)}{3} = 0 \\ \frac{\sum_{i=7}^9 e_i}{3} &= \frac{\sum_{i=7}^9 (\text{Profit}_i - b - m \times 2.00)}{3} = 0\end{aligned}$$

This again gives us three equations with two unknowns, which generally cannot be solved. Consequently, we must think of an alternative way of defining what makes a line best describe the data. To do so, we can interpret it in terms of the residuals, as we did in the dichotomous treatment case. The two intuitive criteria we established by making visual comparisons of lines (Line A vs. Line B and Line C vs. Line D) were:

1. The regression line should not generally tend to overshoot or undershoot the data.
2. The regression line's tendency to over- or undershoot the data should not depend on the level of the treatment (e.g., price).

If we translate these criteria in terms of the residuals, we have:

1. The residuals for all data points average to zero.
2. The size of the residuals is not correlated with the treatment level.

Let's now express those two criteria in equation form. For our Price/Profit example, this gives us:

$$\frac{\sum_{i=1}^9 e_i}{9} = \frac{\sum_{i=1}^9 (\text{Profit}_i - b - m \times \text{Price}_i)}{9} = 0$$

$$\frac{\sum_{i=1}^9 e_i}{9} = \frac{\sum_{i=1}^9 (\text{Profit}_i - b - m \times \text{Price}_i) \times \text{Price}_i}{9} = 0$$

The first equation ensures our residuals average zero across all observations, and the second equation ensures the size of the residuals is not related to the Price level. We now have two equations and two unknowns (m and b), which will generally give us a single solution for our slope and intercept. Solving these two equations for our Price/Profit example yields:

$$m = -48.33$$

$$b = 258.06$$

Therefore, the line that best fits the data, where “best” implies residuals that average zero and are not correlated with the treatment, is:

$$\text{Profit} = 258.06 - 48.33 \times \text{Price}$$

Let's now generalize the criteria we established for our Price/Profit example to any multi-level treatment. To do so, define X as the treatment, which can take on multiple (2 or more) levels, and define Y as the outcome. Then, for a sample of size N , the **simple regression line** is defined as:

simple regression line The slope is the sample covariance of the treatment and outcome divided by the sample variance of the treatment. The intercept is the mean value of the outcome minus the slope times the mean value of the treatment.

$$Y = b + m \times X$$

where b and m are the solution to:

$$\begin{aligned}\frac{\sum_{i=1}^N e_i}{N} &= \frac{\sum_{i=1}^N (Y_i - b - m \times X_i)}{N} = 0 \\ \frac{\sum_{i=1}^N e_i \times X_i}{N} &= \frac{\sum_{i=1}^N (Y_i - b - m \times X_i) \times X_i}{N} = 0\end{aligned}$$

Generically solving for m and b using these two equations yields the following formulas for the slope and intercept of a simple regression line:

$$m = \frac{s\text{Cov}(X,Y)}{s\text{Var}(X)}$$
$$b = \bar{Y} - m \times \bar{X}$$

where:

$$\bar{Y} = \frac{1}{N} \sum_{i=1}^N Y_i$$
$$\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i$$
$$s\text{Cov}(X, Y) = \frac{\sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})}{N-1}$$
$$s\text{Var}(X) = \frac{\sum_{i=1}^N (X_i - \bar{X})^2}{N-1}$$

In words, for a simple regression line, the slope is the sample covariance of the treatment and outcome divided by the sample variance of the treatment. The intercept is the mean value of the outcome minus the slope times the mean value of the treatment.

130

For those less statistically inclined, these formulas might seem a bit cryptic. However, our intent in presenting them is not to imply they should be committed to memory. Rather, they concretely illustrate a simple point: our two equations lead to formulas for the slope and intercept that require nothing more than plugging in summary statistics from our data sample, the four listed previously. In practice, a computer will make these calculations for us, but it is important to recognize and understand at a basic level what the computer is solving and why it makes sense. [Demonstration Problem 5.3](#) provides an opportunity for you to see how to make these calculations directly.

5.3

Demonstration Problem

You are an analyst for FlowersNY, which sells flowers via its website to 18 local

markets across upstate New York. Your firm allocates its advertising budget toward local television advertising in those markets. You've been internally keeping data on your advertising expenditures and the number and location of visits to your website over the past three months. For each location, you have data on the aggregate advertising expenditure and number of visits over the past three months, as displayed in [Table 5.7](#).

Solve for the simple regression line that has website visits as a function of advertising expenditure.

TABLE 5.7 Data on Website Visits and Ad Expenditure for 18 Locations

LOCATION NUMBER	WEBSITE VISITS	AD EXPENDITURE
1	617	32284
2	786	22657
3	493	25024
4	737	41907
5	683	21996
6	622	49047
7	669	26111
8	632	35375
9	1062	52335
10	576	55240
11	992	55269
12	702	42959
13	728	28780
14	229	13796
15	697	39622
16	810	59066
17	517	28398

Answer:

We know for the simple regression line that the formula for the slope is:

$$m = \frac{s\text{Cov}(\text{Ad Exp}, \text{Visits})}{s\text{Var}(\text{Ad Exp})}$$

Plugging in for these values, we get $m = 0.00836$. We also know that $b = \bar{\text{Visits}} - m \times \bar{\text{Ad Exp}}$. Plugging in for these values, we get $b = 369.4$. Therefore, the simple regression line for these data for Visits as a function of Ad Expenditure is: $\text{Visits} = 369.4 + 0.00836 \times \text{Ad Exp}$.

The simple regression line is the line we use to best describe the data for any single treatment. While we solved for the simple regression line in the context of a multi-level treatment, it applies to a dichotomous treatment as well. This is because the regression line for a dichotomous treatment is a special case of the simple regression line. The mathematics that prove this are straightforward but not particularly enlightening. However, to illustrate this point, let's revisit the dichotomous version of the Price/Profit example, and solve for the slope and intercept using the formulas for the simple regression line. We recreate the data for the dichotomous example in [Table 5.8](#).

TABLE 5.8 Price and Profits for Jill's Corn with Two Price Levels

PRICE	PROFITS
\$1.00	\$240
\$1.00	\$200
\$1.00	\$185
\$1.50	\$205
\$1.50	\$170

\$1.50

\$190

Using the formulas for variance and covariance, we have $s\text{Cov}(\text{Profit}, \text{Price}) = -3$ and $s\text{Var}(\text{Price}) = 0.075$. We also have $\bar{\text{Price}} = 1.25$ and $\bar{\text{Profit}} = 198.33$. Plugging these into our formulas for slope and intercept, we have $m = -3/0.075 = -40$, and $b = 198.33 - (-40) \times 1.25 = 248.33$. Note that these are exactly the same slope and intercept we arrived at before.

Looking ahead, when solving for the slope and intercept of the regression line for a single treatment, we will generally use the formulas we established for the simple regression line, whether the treatment is dichotomous or multi-level. However, the dichotomous case builds a useful bridge back to the fundamentals of the scientific method, and it will prove useful in illustrating the difference between correlation and causality, a topic we discuss in detail in [Chapter 6](#).

We summarize the reasoning for the simple regression line in [Reasoning Box 5.2](#).

132

REASONING BOX 5.2

THE SIMPLE REGRESSION LINE

For the case of any single treatment (dichotomous or multi-level), let $Y = b + mX$ be a line describing a given dataset with N observations, where Y is the outcome and X is the treatment. Define a line as best describing the data if:

1. The residuals for all data points average to zero. Mathematically:

$$\frac{\sum_{i=1}^N e_i}{N} = \frac{\sum_{i=1}^N (Y_i - b - m \times X_i)}{N} = 0$$

2. The size of the residuals is not correlated with the treatment level.

Mathematically:

$$\frac{\sum_{i=1}^N e_i \times X_i}{N} = \frac{\sum_{i=1}^N (Y_i - b - m \times X_i) \times X_i}{N} = 0$$

Then, the slope and intercept for this line are:

$$m = \frac{s\text{Cov}(X,Y)}{s\text{Var}(X)}$$

$$b = \bar{Y} - m \times \bar{X}$$

Consequently, we can write the line best describing the data as:

$$Y_i = \left(\bar{Y} - \frac{s\text{Cov}(X,Y)}{s\text{Var}(X)} \times \bar{X} \right) + \frac{s\text{Cov}(X,Y)}{s\text{Var}(X)} \times X_i$$

This is defined as the simple regression line.

COMMUNICATING DATA 5.1

REGRESSION LINE ORIGINS

A natural question after defining regression lines is to ask: "Why is it called regression?" The origin of the regression moniker for these lines comes from Sir Francis Galton. He took on the task of collecting data on parents' and their children's heights. He then plotted these data and attempted to fit a line through them to describe the relationship between these measures, where the child's height plays the role of the Outcome and the parent's height plays the role of the Treatment. For simplicity, we can consider a version of this exercise where we relate a son's height to his father's height as Son = $b + m \times$ Father. When fitting a line to such data, Galton found a slope that was less than one. To be concrete, let's suppose he found a slope of 0.8. What can we infer from this measurement?

Suppose the average height of a man is 70 inches. Our line suggests that if a father is five inches above average (75 inches), then his son tends to retain only 80% of this extra height along our line—that is, he is four inches above

average (74 inches). And, if a father is five inches below average (65 inches), again his son tends to retain only 80% of this lost height along our line—he is four inches below average (66 inches). In sum, this line implies a regression to the mean for height; along the line, tall fathers have tall, but less tall sons, and short fathers have short, but less short sons.

It was from this early application that the regression line was born, despite the fact that Galton's method of fitting the line did not exactly match the method we described in [Reasoning Box 5.2](#). The next natural question is to ask whether the regression line serves as an appropriate prediction tool for the general population of fathers and sons, a topic we discuss in detail in [Chapter 6](#).

Sample Moments and Least Squares

LO 5.4 Distinguish the use of sample moment equations from estimation via least squares.

The two criteria we established toward constructing a simple regression line centered on the residuals and the product of the residuals and the treatment. We imposed conditions on $\frac{1}{N} \sum_{i=1}^N e_i$ and $\frac{1}{N} \sum_{i=1}^N e_i \times X_i$, setting both equal to zero. These two expressions are examples of sample moments. Broadly speaking, a **sample moment** is the mean of a function of a random variable(s) for a given sample. For example, if we have a sample of size 20 that contains information on salaries, $\frac{1}{20} \sum_{i=1}^{20} \text{Salary}_i^3$ is a sample moment, where Salary_i is the random variable and the function is generically defined as $f(a) = a^3$.

sample moment The mean of a function of a random variable(s) for a given sample.

In short, the approach we used to solve for a line that best describes a

given dataset with a single treatment involved setting two sample moments equal to zero. While we have presented intuitive arguments why imposing these two conditions will yield a “good” line, there are other perspectives concerning how one should determine what line is “best.” By far, the most utilized is simply known as the method of linear least squares, or ordinary least squares (OLS).

To determine a line that best describes the data, OLS establishes an objective and then deems the line that accomplishes the objective as the line that best describes the data. As with our sample moments, OLS centers on the residuals. OLS defines its objective to be the minimization of the sum of the squared residuals. Formally, **ordinary least squares** solves:

ordinary least squares The process of solving for the slope and intercept that minimize the sum of the squared residuals.

$$\text{Min}_{b,m} \sum_{i=1}^N e_i^2 = \sum_{i=1}^N (Y_i - b - m \times X_i)^2$$

Ordinary least squares says that the slope and intercept that solve the above minimization problem correspond to the line that best describes the data.

As was the case with our sample moments, OLS has a strong intuitive motivation. Recall that the residual is the difference between the observed outcome and the corresponding point on the regression line for a given observation. OLS seeks the line that makes these differences small, since small residuals imply the line is generally close to the observed data points. It does so by establishing an **objective function**, a function we ultimately wish to maximize or minimize (minimize in this case). For ordinary least squares, the objective function is the sum of squared residuals ($\sum_{i=1}^N e_i^2$)—a function that is increasing in the residuals—and then the OLS method attempts to find the slope and intercept that minimize it, resulting in small residuals.

objective function A function ultimately wished to be maximized or minimized.

While there are intuitive arguments for using the sum of squared

residuals as our objective function when seeking a line that best describes the data, it is not the only plausible choice. We could instead use the sum of the absolute value of the residuals as our objective function, i.e., $\sum_{i=1}^N |e_i|$, and solve for the slope and intercept that minimize it. This objective function is used regularly to construct lines to describe datasets, and this method is known as **least absolute deviations (LAD)**. Other methods with alternative objective functions also exist, and each generally will produce a different line to describe the data.

least absolute deviations (LAD) Use the sum of the absolute value of the residuals as the objective function and solve for the slope and intercept that minimize it.

134

Why then is OLS the dominant method in practice? There are some technical reasons, e.g., it always has a unique solution while LAD may not. However, an appealing property of OLS is that it exactly lines up with the intuition that led us to construct the simple regression line using only intuitive arguments. Specifically, the solution to the

COMMUNICATING DATA 5.2

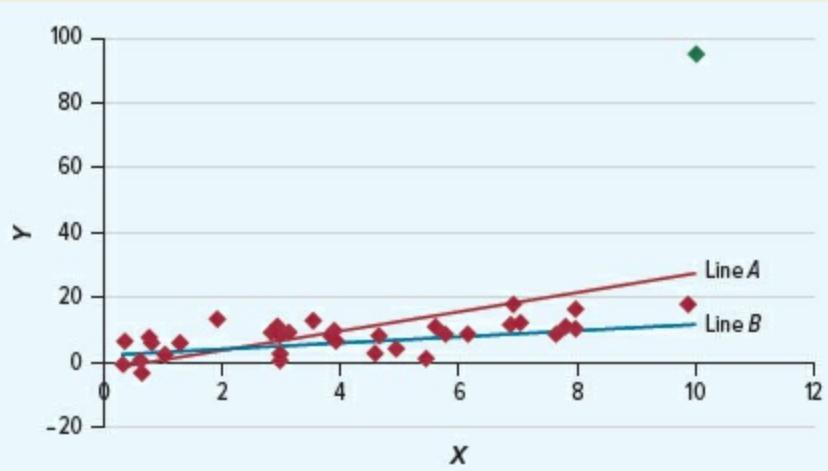
LEAST SQUARES VS. LEAST ABSOLUTE DEVIATIONS

Least squares is often taken for granted as THE way to fit a function through a dataset. However, there are other ways to do it, including *least absolute deviations*. This is important to understand, because it's a reminder that using least squares to solve for our function describing the data is a choice, with consequences for the estimates that differ from other alternatives. For example, let's compare least squares to least absolute deviations. Rather than doing so with technical equations, let's consider a conceptual comparison for the simple case of using a line to describe a dataset.

Suppose our dataset looks like what is shown in [Figure 5.10](#), where there

appears to be a clear, positive relationship between Y and X . In the figure, we present two candidate lines to describe the data. Can you guess which comes from OLS and which from LAD?

FIGURE 5.10 OLS vs. LAD for Describing a Dataset



The key lies in the clear outlier, the point that looks very unlike the others, with a Y value near 100 (colored green). This outlier is going to inevitably generate a very large residual (e), which has a big negative impact on the objective function for both OLS (sum of squared residuals) and LAD (sum of absolute value of residuals), since our objective is to minimize. However, it's easy to see the impact will be greater in OLS, since this large value is squared (e.g., if it is 80, it contributes 6400 in OLS and just 80 in LAD to the objective function). Therefore, the line we solve for in OLS will try to "accommodate" this outlier, by getting closer to it, more than the line we solve for in LAD. It's as if the outlier is pulling the line toward it with more force in OLS than in LAD.

Consequently, the answer to our question is that Line A is from OLS and Line B is from LAD, as evidenced by the fact that Line A gets closer to the outlier. This simple example illustrates a key difference between OLS and LAD: OLS is more sensitive to outliers than LAD.

OLS problem results in *exactly the same* sample moment equations we derived directly in the previous section, and hence the same solution for the slope and intercept. Solving $\text{Min}_{b,m} \sum_{i=1}^N e_i^2$ results in the same b and m you get when you solve $\frac{1}{N} \sum_{i=1}^N e_i = 0$ and $\frac{1}{N} \sum_{i=1}^N e_i \times X_i = 0$. Whether we specify that we are using OLS or imposing that these two sample moments are zero, we end up with the same line, the simple regression line.

$$m = \frac{s\text{Cov}(X,Y)}{s\text{Var}(X)}$$

$$b = \bar{Y} - m \times \bar{X}$$

In practice, when an analyst produces the simple regression line to describe a dataset, she will state that she used OLS to do so. Of course, this is correct, since OLS does produce the simple regression line. However, there are good reasons to conceptualize the simple regression line (and multiple regression line, discussed below) in terms of the sample moments rather than the minimization of the sum of squared residuals. First, sample moments give you the conditions that ultimately produce the slope and intercept directly, rather than just specifying the minimization problem to be solved (as in OLS). This focuses attention on the criteria we are imposing on the residuals (how the line relates to the observed data points) to produce the line, which are not readily obvious when just considering an objective function. Second, as will be clear in [Chapter 6](#), thinking in terms of the sample moments facilitates our ability to assess whether or not our OLS line describes a causal relationship between the treatment and outcome (are we getting an unbiased estimate of the treatment effect?). Lastly, thinking in terms of sample moments simplifies the process of extending regression analysis into the case of multiple treatments, to which we now turn.

Regression for Multiple Treatments

SINGLE VS. MULTIPLE TREATMENTS

LO 5.5 Distinguish regression equations for single and multiple treatments.

To this point, our discussion has focused on finding a line to describe data that involve a single treatment, resulting in the formula for the simple regression line. However, there are many instances in business and beyond where more than one treatment is involved.

As a simple example, let's revisit the medical field, where we are interested in the relationship between two separate drugs and individuals' health outcomes. The health outcome of interest is a person's cholesterol level, measured in milligrams (mg), and we have two drugs, Drug A and Drug B, also measured in mg. We have data on the cholesterol levels, and dosages taken for each drug, for 15 individuals at a given time. Let the cholesterol level be the measurement taken by a doctor on January 1, and the dosages be the average dosage per week over the prior six weeks. See [Table 5.9](#).

136

TABLE 5.9 Cholesterol Level and Drug Dosages for 15 Individuals

INDIVIDUAL	CHOLESTEROL	DRUG A DOSAGE	DRUG B DOSAGE
1	207	35	2
2	224	31	8
3	192	42	44
4	163	41	8
5	186	27	34
6	230	48	1
7	222	5	35

8	218	15	42
9	182	34	43
10	224	33	24
11	236	15	41
12	224	20	12
13	177	49	16
14	182	39	46
15	181	47	13

With these data, we could consider the relationship between cholesterol and each drug individually, and estimate the simple regression line for each, arriving at:

$$\text{Cholesterol} = 235.17 - 0.997 \times \text{Drug A}$$

$$\text{Cholesterol} = 205.83 - 0.107 \times \text{Drug B}$$

These simple regression lines are effective at showing us the pairwise relationships between each treatment and the outcome. However, it can be beneficial to describe the relationship between both treatments and the outcome with a single expression, important when we utilize regression models to make predictions.

When we had a single treatment, we used a line to associate a value for the outcome with any given value for the treatment. For example, when we restrict ourselves to just Drug A as the single treatment, then the corresponding simple regression line assigns a value for cholesterol to each dosage level of Drug A. When we have two treatments, we want an expression that associates a value for the outcome with any given combination of values for the treatments. For our example, we want a value for cholesterol to be associated with any combination of dosages for Drug A

and Drug B. This requires something other than a line—it requires a plane. In general, the expression for a plane can be written as $Y = b + m_1X_1 + m_2X_2$. For our example, this translates into: Cholesterol = $b + m_1$ Drug A + m_2 Drug B. Notice two things: It is a natural extension of a line for three dimensions, and it allows us to associate a value for Cholesterol for any combination of dosages for Drug A and Drug B.

137

As we are seeking a plane to describe the data, the question is what plane best describes the data? In principle, we could extend the intuition we used when visually comparing lines (Line A vs. Line B) for a single multi-level treatment to visually compare planes. However, such an approach can become complicated quickly, thus compromising any intuition it may try to provide. Instead, we turn to the residuals and extend the approach presented there from the case of a single treatment to that of two treatments. For our cholesterol example, we write each cholesterol outcome as follows:

$$\text{Cholesterol}_i = b + m_1 \text{ Drug A}_i + m_2 \text{ Drug B}_i + e_i$$

As before, e_i is the residual, the difference between the observed outcome and the corresponding point on the plane for a given observation.

Recall that our criteria for the residuals to establish a “best” line were:

1. The residuals for all data points average to zero.
2. The size of the residuals is not correlated with the treatment level.

Conveniently, these criteria very simply extend to our criteria for a best plane. We need only make a minor change to the second criterion to arrive at these modified criteria for a plane to best describe the data:

1. The residuals for all data points average to zero.
2. The size of the residuals is not correlated with the treatment level *for any treatment*.

Note that all we have done is added the stipulation that residuals are not correlated with the treatment level for any treatment. This means the residuals are not correlated with the dosage level for Drug A or Drug B. This naturally and easily extends to any number of treatments; if there are K treatments, then our second criterion implies the residuals are not correlated with the treatment level for Treatment 1, Treatment 2, ..., and Treatment K .

As we did with a single treatment, let's now express these criteria in equation form. For our cholesterol example, this gives us:

$$\frac{\sum_{i=1}^{15} e_i}{15} = \frac{\sum_{i=1}^{15} (\text{Cholesterol}_i - b - m_1 \times \text{Drug A}_i - m_2 \times \text{Drug B}_i)}{15} = 0$$

$$\frac{\sum_{i=1}^{15} e_i \times \text{Drug A}_i}{15} = \frac{\sum_{i=1}^{15} (\text{Cholesterol}_i - b - m_1 \times \text{Drug A}_i - m_2 \times \text{Drug B}_i) \times \text{Drug A}_i}{15} = 0$$

$$\frac{\sum_{i=1}^{15} e_i \times \text{Drug B}_i}{15} = \frac{\sum_{i=1}^{15} (\text{Cholesterol}_i - b - m_1 \times \text{Drug A}_i - m_2 \times \text{Drug B}_i) \times \text{Drug B}_i}{15} = 0$$

We could use these three equations to derive formulas for b , m_1 , and m_2 , but unfortunately, these formulas are a bit more complex compared to those for the simple regression line. However, we can use the Regression capability for any statistical software to have a computer produce the solution for b , m_1 , and m_2 using these three equations. For example, if we regress Cholesterol on Drug A and Drug B using the data in [Table 5.9](#) in Excel, the output will look as in [Table 5.10](#).

[Table 5.10](#) contains a substantial amount of information, much of which we will discuss throughout this book. For our cholesterol example, we need only focus on the highlighted

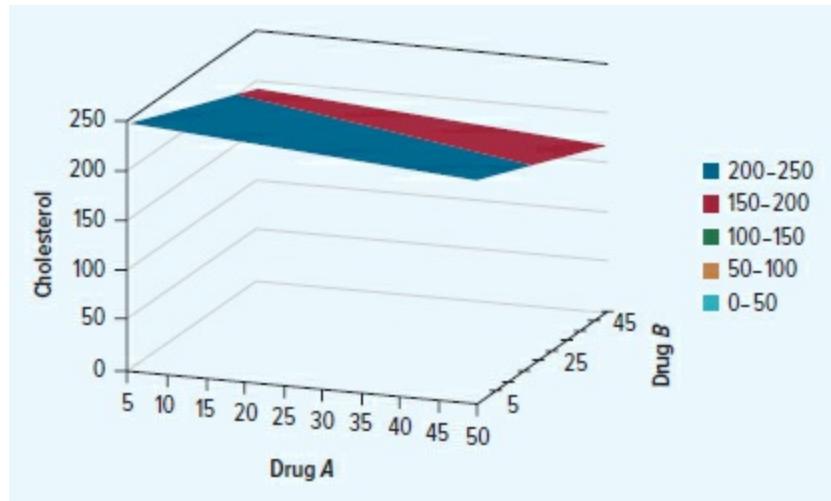
TABLE 5.10 Regression Output in Excel for Cholesterol Regressed on Drug A and Drug B

Regression Statistics				
Multiple R	0.65475068			
R Square	0.428698454			

Adjusted R Square	0.333481529			
Standard Error	19.26342124			
Observations	15			
ANOVA				
	df	SS	MS	F
Regression	2	3341.447227	1670.723613	4.5023346
Residual	12	4452.952773	371.0793978	
Total	14	7794.4		
	Coefficients	Standard Error	t Stat	P-value
Intercept	256.2033683	19.17165912	13.36365135	1.44509E-05
Drug A	-1.258763673	0.422324237	-2.980562239	0.0114737
Drug B	-0.513781026	0.336328316	-1.527617516	0.152527

cells. Here, we have the values for b (256.20), m_1 (-1.259), and m_2 (-0.514) that the computer generated when solving the three previous equations. Hence, we have the equation for our regression plane: Cholesterol = 256.20 - 1.259 × Drug A - 0.514 × Drug B. We graph this equation in [Figure 5.11](#).

FIGURE 5.11 Regression Plane for Cholesterol Regressed on Drug A and Drug B



139

MULTIPLE REGRESSION

LO 5.6 Describe a dataset with multiple treatments using multiple regression.

For the case of a single treatment, we used a regression line to describe the relationship between the outcome and treatment for a given dataset. We extended this approach to the case of two treatments, where we used a plane to describe the relationship between the outcome and our two treatments for a given dataset. Now, we extend the analysis to any number of treatments.

Suppose we have K treatments, where $K \geq 2$. We extend our cholesterol example to the case where there are K drugs, now labeled as Drug 1, Drug 2, ..., Drug K , rather than just two. Just as we did with one or two treatments, we want an expression that associates a value for the outcome with any given combination of values for the treatments. For our extended cholesterol example, we want a value for cholesterol to be associated with any combination of dosages for Drug 1, Drug 2, ... and Drug K .

When we have more than two treatments, accomplishing this task requires us to solve for a hyperplane. While this may sound technical, it is actually a simple extension of our plane for the case of two treatments. We

can write the expression for a hyperplane when there are K treatments as $Y = b + m_1X_1 + m_2X_2 + \dots + m_KX_K$. For our example, this translates into: Cholesterol = $b + m_1$ Drug 1 + m_2 Drug 2 + ... + m_K Drug K . As the regression plane was a simple extension of the regression line, the regression hyperplane is a natural extension of the regression plane. Notice that this equation again allows us to associate a value for Cholesterol for any combination of dosages for Drug 1 through Drug K .

As we did with one and two treatments, we'd like to determine the hyperplane that best describes the data for K treatments. We can again express our observations as outcomes written in terms of the treatments and a residual. For our extended cholesterol example, we can write each cholesterol outcome as follows: $\text{Cholesterol}_i = b + m_1 \text{ Drug } 1_i + m_2 \text{ Drug } 2_i + \dots + m_K \text{ Drug } K_i + e_i$. It is for this general case of K treatments that we can see one of the benefits of thinking in terms of sample moments. In fact, we needn't make any change to the criteria we used to find the "best" regression plane when we extend to the more general case of a hyperplane. The criteria remain as:

1. The residuals for all data points average to zero.
2. The size of the residuals is not correlated with the treatment level *for any treatment*.

For the case of K treatments, the second criterion means that the residuals are not correlated with the dosage level for any of the drugs, Drug 1 *through* Drug K .

Then, for a sample of size N with K treatments, the associated equations for these criteria are:

$$\frac{\sum_{i=1}^N e_i}{N} = \frac{\sum_{i=1}^N (\text{Cholesterol}_i - b - m_1 \times \text{Drug } 1_i - \dots - m_K \times \text{Drug } K_i)}{N} = 0$$

$$\frac{\sum_{i=1}^N e_i \times \text{Drug } 1_i}{N} = \frac{\sum_{i=1}^N (\text{Cholesterol}_i - b - m_1 \times \text{Drug } 1_i - \dots - m_K \times \text{Drug } K_i) \times \text{Drug } 1_i}{N} = 0$$

$$\frac{\sum_{i=1}^N e_i \times \text{Drug } K_i}{N} = \frac{\sum_{i=1}^N (\text{Cholesterol}_i - b - m_1 \times \text{Drug } 1_i - \dots - m_K \times \text{Drug } K_i) \times \text{Drug } K_i}{N} = 0$$

As with the case of two treatments, we could use these equations to derive formulas for b , m_1 , m_2 , ..., m_K , but since these formulas are a bit messy, we generally use a computer to solve them (in Excel as illustrated in [Demonstration Problem 5.4](#)).

In this section, we have extended the case of one treatment, for which we utilize the simple regression line, to the case of multiple treatments. Any (hyper)plane that we use to describe the data can be characterized as a multiple regression (hyper)plane. As was the case for a single treatment, the regression (hyper)plane that satisfies the given sample moment equations is the same as the regression (hyper)plane that you would solve for using OLS. Consequently, it is convention to call this (hyper)plane the OLS multiple regression (hyper)plane. However, we seldom use this specific expression in practice. Instead, we typically refer to the process that generates this (hyper)plane—solving the above equations involving sample moments, or equivalently solving OLS—as “OLS multiple regression.”

Note however that, although there are many other ways to solve for a function that best describes the data (e.g., LAD), the expression **multiple regression** by itself is understood to imply the use of OLS (or equivalently, the sample moment equations). Analogously, the process that produces the simple regression line for a single treatment (solving the two sample moment equations) is called **simple regression**.

multiple regression Solving for a function that best describes the data that implies the use of OLS (or equivalently, the sample moment equations).

simple regression The process that produces the simple regression line for a single treatment.

REASONING BOX 5.3

MULTIPLE REGRESSION

For the case of multiple treatments, let $Y = b + m_1X_1 + \dots + m_KX_K$ be a line describing a given dataset with N observations and K treatments. Here, Y is the outcome and X_1, \dots, X_K describe the K different treatments. Define a line as best describing the data if:

1. The residuals for all data points average to zero. Mathematically:

$$\frac{\sum_{i=1}^N e_i}{N} = \frac{\sum_{i=1}^N (Y_i - b - m_1X_{1i} - \dots - m_KX_{Ki})}{N} = 0$$

2. The size of the residuals is not correlated with the treatment level for any treatment. Mathematically:

$$\begin{aligned}\frac{\sum_{i=1}^N e_i \times X_{1i}}{N} &= \frac{\sum_{i=1}^N (Y_i - b - m_1X_{1i} - \dots - m_KX_{Ki}) \times X_{1i}}{N} = 0 \\ \frac{\sum_{i=1}^N e_i \times X_{Ki}}{N} &= \frac{\sum_{i=1}^N (Y_i - b - m_1X_{1i} - \dots - m_KX_{Ki}) \times X_{Ki}}{N} = 0\end{aligned}$$

The process of solving this system of equations for b, m_1, \dots, m_K (typically done using a computer) is technically called *OLS multiple regression* (and often shortened to *multiple regression*). It results in a multiple regression plane (when $K = 2$) or multiple regression hyperplane (when $K > 2$).

5.4 Demonstration Problem

Suppose you are working with the dataset contained in [Table 5.11](#), composed of an outcome (Y) and five treatments (X_1, \dots, X_5). You are seeking to describe these data with the following function: $Y = b + m_1X_1 + m_2X_2 + m_3X_3 + m_4X_4 + m_5X_5$. Use (OLS) multiple regression to find the values for b, m_1, \dots, m_5 .

TABLE 5.11 Data on an Outcome Y and Five Treatments, X_1-X_5

Y	X_1	X_2	X_3	X_4	X_5
423	167	294	44	107	234
474	179	251	103	97	206
463	84	283	204	234	176
289	79	10	281	213	94
834	191	217	230	30	180
724	175	126	228	89	45
315	165	176	105	182	253
161	39	39	128	147	2
488	26	215	160	66	292
494	170	221	88	18	78
922	175	299	229	33	22
1044	49	294	271	173	7
103	173	44	83	267	135
776	102	253	257	148	54

Answer:

Using the Regression tool in Excel, and regressing Y on the block of data X_1 through X_5 , produces the values in [Table 5.12](#). Therefore, our estimated regression equation, using OLS (i.e., solving six sample moment equations) is:

$$Y = -35.65 + 0.67 \times X_1 + 1.59 \times X_2 + 2.06 \times X_3 - 0.84 \times X_4 - 0.54 \times X_5.$$

TABLE 5.12 Estimates from Multiple Regression of Y on X_1-X_5

INTERCEPT	-35.65232844
X_1	0.670315371
X_2	1.588934631
X_3	2.059616134
X_4	-0.83639557

What Makes Regression Linear?

LO 5.7 Explain the difference between linear regression and a regression line.

Throughout this chapter, we have presented and discussed examples of regression lines and regression planes and hyperplanes. However, despite the title of the chapter including the term “linear regression,” we have yet to use that term in our discussion. This omission until now was, of course, deliberate, as the term “linear regression” can be the cause of confusion when utilizing and explaining these models. In particular, it is not uncommon for “linear regression” to be used interchangeably with (simple) “regression line.” After all, both terms have the word “regression,” and a version of the word “line” in them, so why are they different?

The processes of simple regression and multiple regression, as we’ve discussed in this chapter, actually both fall under the general category of linear regression. **Linear regression** is the process of fitting a function that is linear in its *parameters* to a given dataset. Technically speaking, this means the parameters all enter the function in the first degree (i.e., they all have exponents of one). Practically speaking, it means we can write the function we use to fit the data generically as:

linear regression The process of fitting a function that is linear in its parameters to a given dataset.

$$Y = b + m_1X_1 + m_2X_2 + \dots + m_KX_K$$

Here, $\{b, m_1, \dots, m_K\}$ are the parameters for this function, and note that they all have an exponent of one.

As a point of contrast, if we used the following function to fit a given dataset, we would not be using linear regression (in fact, this would be nonlinear regression, a topic outside the scope of this book):

$$Y = b + m_1 X_1^{m_2} + m_3^2 X_2$$

In this contrasting example, both m_2 and m_3 enter the equation in a nonlinear way.

Defined this way, it is easy to see that both the simple and multiple regression we've defined fit into the category of linear regression. In fact, unless explicitly specified otherwise, the terms "simple regression" and "multiple regression" implicitly contain the word "linear." That is, "simple regression" implies "simple *linear* regression" and "multiple regression" implies "multiple *linear* regression."

To conclude, we aim to resolve a common confusion with linear regression. The use of linear regression does not at all imply we necessarily will be constructing a line to fit the data. Linear regression is linear in the parameters but not necessarily the treatment(s). For example, consider the following generic multiple linear regression equation for two treatments: $Y = b + m_1 X_1 + m_2 X_2$. Now, suppose we defined $Y = \text{Cholesterol}$, $X_1 = \text{Drug 1}$, and $X_2 = \text{Drug 1}^2$. Here, we technically have two treatments, but the second is just a function of the first. Consequently, although this is a multiple regression equation, we can graph the relationship in a two-dimensional graph (Cholesterol graphed against Drug 1).

The fact that some of the X s in multiple regression can simply be functions of other X s makes linear regression quite versatile. It allows for an unlimited number of possible "shapes" for the relationship between the outcome and any particular treatment. We go into much greater detail about modeling the shape of the relationship between

TABLE 5.13 Data on Cholesterol Level and Dosage of Drug A

CHOLESTEROL	DRUG A
206	50
219	40
286	1
244	9
201	29
216	14
202	44
234	15
216	7
184	34
268	4
193	29
247	3
139	25
203	29
181	37

outcome and treatment in [Chapter 7](#), when we discuss functional form issues. However, we present a simple example here to illustrate how linear regression can produce much more than just lines (and hyperplanes) to describe the relationship between a treatment(s) and outcome.

Consider the data in [Table 5.13](#), where we have information on cholesterol and just a single drug, Drug A. With just a single treatment, we could utilize simple linear regression to model the relationship between Cholesterol and Drug A in the data. Here, we estimate the equation

Cholesterol = $b + m \times$ Drug A. Using this approach, we are applying linear regression—since the model is linear in the parameters—and also estimating a line, since the equation we are estimating is linear in the treatment (Drug A). When we solve for the intercept and slope, we get the estimated relationship of: Cholesterol = 249.02 – 1.47 × Drug A.

Estimating a linear relationship between Cholesterol and Drug A is far from our only option when applying linear regression. For instance, we may want to estimate a quadratic relationship, or a cubic relationship. For the former, we want to fit the equation Cholesterol = $b + m_1$ Drug A + m_2 Drug A², and for the latter, we want to fit the equation Cholesterol = $b + m_1$ Drug A + m_2 Drug A² + m_3 Drug A³. In both cases, we are now utilizing multiple (linear) regression. Each model is linear in the parameters, but now neither is linear in the treatment (Drug A). When we solve for the parameters for these models, we get: Cholesterol = 278.24 – 5.73 × Drug A + 0.09 × Drug A² and Cholesterol = 285.37 – 7.67 × Drug A + 0.19 × Drug A² + 0.001 × Drug A³, respectively.

144

In sum, linear regression encapsulates a wide range of models for the relationship between an outcome and treatment(s). These can involve any number of different treatments, and the outcome being either a linear or nonlinear function of any of these treatments. The only restriction for linear regression is that the equation we fit to the data be linear in the parameters. Fortunately, this requirement typically is met for a large proportion of functions used to fit datasets in business and beyond.

5.5

Demonstration Problem

Which of the following are linear regression models?

1. $Y = b + m_1 \times X_1 - m_2 \times X_2$

2. Sales = $b + m_1 \times \text{Price} + m_2 \times \text{Price}^2$
3. Production = $b \times \text{Labor}^{m_1}$
4. $Y = b + m_1 X_1 + m_2 X_2 + m_3 X_2^2$

Answer:

1. Linear regression
2. Linear regression
3. Not linear regression
4. Linear regression

COMMUNICATING DATA 5.3

REGRESSION FOR RATINGS

Your firm has launched a new yogurt product, called Jogurt, named for its founder (Joe). You are hoping to learn more about what types of consumers prefer this new product, so you offer free samples in grocery stores to interested customers in exchange for some basic consumer information. In particular, you collect information on their rating of the product (on a scale from 1 to 7), their age, and their household size.

You run two regressions. One uses the simple regression line: Rating = $b + m \times \text{Age}$. The other uses multiple regression to find the version of Rating = $b + m_1 \times \text{Age} + m_2 \times \text{HHSIZE}$ that best describes the data. Your estimates yield Rating = $5.1 - 0.03 \times \text{Age}$ and Rating = $5.4 - 0.05 \times \text{Age} + 0.4 \times \text{HHSIZE}$.

With just this information, what can you correctly communicate about your analysis? On the more technical side, you know the estimated parameters for both equations yield the line/plane that best describes the data, where best implies satisfaction of the sample moment equations (i.e., residuals average zero and are uncorrelated with the Treatment(s)/X(s)). On the conceptual side, both describe how your outcome (Rating) moves with your treatment(s) (Age, Age and HHSIZE).

Looking at the first equation, it appears that, in your data, Rating tends to decline (slightly) with Age. From the second equation, it appears that, when moving up the Age dimension, Rating tends to slightly decline while at the same time, Rating tends to increase when moving up the Household Size dimension. The sizes of the estimates give a sense of how “steep” the data are in any given direction (Age, Household Size).

To say more, we need to make some assumptions, which we discuss in detail in the next chapter.

RISING TO THE **data**CHALLENGE

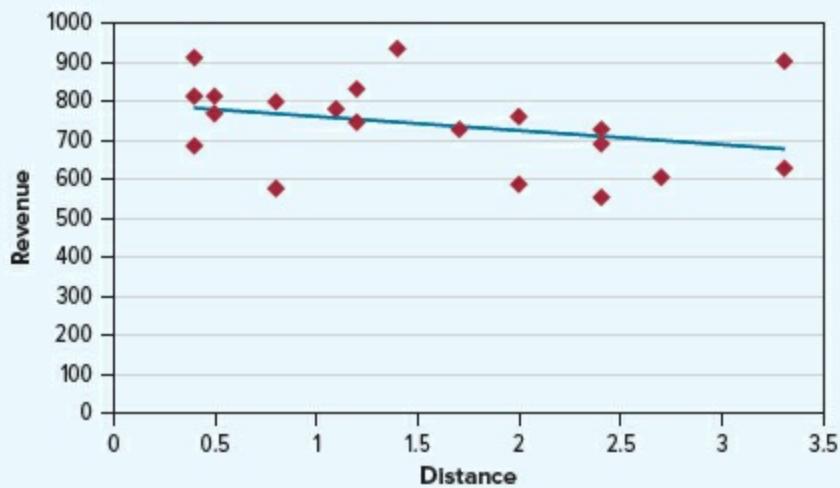
Where to Park Your Truck?

Let’s return to the Data Challenge posed at the start of the chapter: where to park your food truck in a large college town. We can describe the relationship between revenue and distance from the university by using a simple regression line. In this case, the line’s equation is $\text{Revenue} = b + m \times \text{Distance}$, and we can define each individual data point as $\text{Revenue}_i = b + m \times \text{Distance}_i + e_i$, where e_i —the residual—is the difference between the observed Revenue and the corresponding point on the regression line. If we define a line as best describing the data if the residuals are zero on average and are uncorrelated with Distance, then the line that best describes these data has equation:

$$\text{Revenue} = 803.97 - 37.52 \times \text{Distance}$$

We plot this regression line along with the data points in [Figure 5.12](#).

FIGURE 5.12 Scatterplot of Revenue and Distance along with Regression Line



Here we see that the line best describing the data has Revenue declining with Distance but at a relatively modest rate.

SUMMARY

This chapter presented regression analysis as a descriptive tool for a given dataset. It began with the simplest case, describing data consisting of an outcome and a dichotomous treatment. It then expanded into multi-level treatments and multiple treatments. In doing so, we started by detailing the regression line, and then explained how regression can use many other functions besides a line (e.g., quadratic, cubic) to describe a dataset. All of the regression analysis discussed in this chapter falls into the category of linear regression, where the functions used to describe the data are linear in their parameters, but not necessarily the treatments (allowing the use of nonlinear functions like quadratics and cubics).

146

It is important to highlight that every discussion and every example in this chapter has dealt with regression analysis as a descriptive tool. None of the functions we fit to the data were claimed to measure treatment effects; they simply describe how the outcome “moved” with the treatment(s) for the given datasets. Regression analysis can be highly useful toward measuring treatment effects, but it takes more than simply solving sample moment equations; it

requires a line of reasoning. We turn to this topic in the next chapter.

KEY TERMS AND CONCEPTS

dichotomous treatment

least absolute deviations (LAD)

linear regression

multi-level treatment

multiple regression

objective function

ordinary least squares (OLS)

regression analysis

regression line for a dichotomous treatment

residual

sample moment

simple regression

simple regression line

CONCEPTUAL QUESTIONS connect

1. Your software firm just finished developing an upgrade for one of its popular applications. It began by offering the product in select markets and then recorded sales in markets with the old version (V1.0) and the new version (V2.0). Thus far, mean sales per 100,000 (annualized) for V1.0 were 125, and for the new version were 187. Solve for the regression line for a dichotomous treatment that describes the relationship between sales and the version of your application. (LO1)
2. Suppose you have given your research analyst data on an outcome (Y) and treatment (X) and asked her to estimate the simple regression line. She does this and produces the following table, containing your data

points for Y , X , and the residuals generated from the regression line she estimated. Unfortunately, there are some missing values in the table, and she did not give you the equation of the regression line. However, with your knowledge of regression, you are able to determine the regression line equation with just this information. What is the equation of the regression line? (LO2)

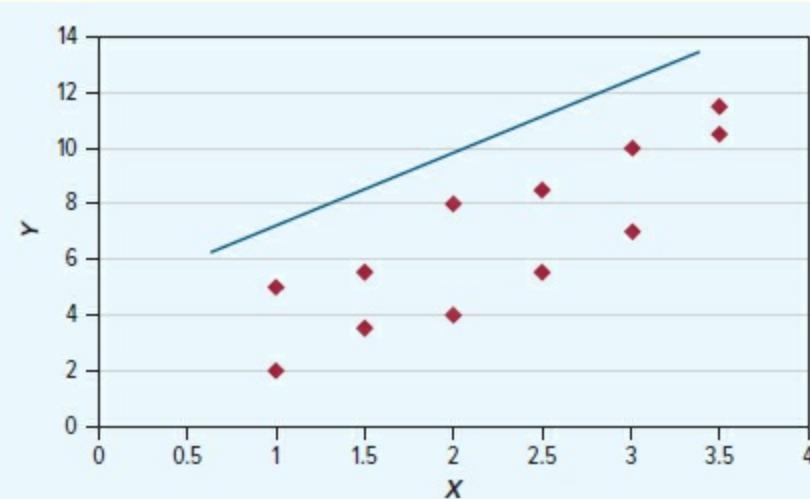
Y	X	RESIDUAL
27		-2.645
16		0.376
32		-2.903
19		-0.129
25		-1.140
37	5	-1.408
41	8	

147

-
3. Explain the difference between linear regression and a regression line. (LO7)
 4. You would like to describe the relationship between customers' ratings of your product and their income and education, using the regression equation: Rating = $b + m_1 \times \text{Income} + m_2 \times \text{Education}$. What are the sample moment equations you would use to solve for b , m_1 , and m_2 , expressed using only the terms: Rating, Income, Education, b , m_1 , and m_2 and N (the number of observations)? (LO6)
 5. Which of the following are linear regression models? (LO7)
 - a. $\text{Salary} = b + \text{Tenure}^{m_1}$
 - b. $\text{Salary} = b + m_1 \text{Tenure}^2$
 - c. $\text{Production} = m_1 + \sqrt{\text{Labor}}$
 - d. $\text{Production} = m_1 + \text{Labor} \times m_2 \sqrt{\text{Capital}}$
 6. Suppose you are trying to estimate the regression equation $Y = b + m \times X$ for a dataset containing information on Y and X . You decide to use OLS to

get your estimates. Compare and contrast the estimates you get using OLS to what you would have gotten, had you used the sample moment equations. (LO4)

7. Explain intuitively why the line presented in the following figure cannot be the corresponding regression line for the data points in the figure. (LO3)
8. Suppose $s\text{Cov}(X, Y) < 0$. If we regress Y on X , will the slope of the regression line be positive, zero, negative, or is it impossible to tell? Explain. (LO3)



9. In general, if you solve for a regression line using least absolute deviations (LAD), will this produce the same line as the one you would get solving the sample moment equations? Why or why not? (LO4)
10. Suppose you have data on three separate variables: Y , X , and Z . Suppose further that you have estimated the following two regression equations:

$$Y = 15 - 2.7 \times X$$

$$Y = 21 + 3.4 \times Z$$

Rather than have two separate simple regressions, you want a single regression of Y on X and Z , written as $Y = b + m_1X + m_2Z$. If you estimate b , m_1 , and m_2 using the same data you used to get your two simple regressions, what will the values of b , m_1 and m_2 be? Do you need further

information to solve for any of those three values? (LO5)

148

- 11.** Suppose you are estimating the following regression equation: $Y = b + m_1X_1 + \dots + m_KX_K$ (LO5)
- How many sample moment equations must you solve if:
 - $K = 1$
 - $K = 5$
 - $K = 30$
 - Supposing $K = 30$, summarize in two sentences the criteria that the associated sample moment conditions used to solve the regression equation satisfy.

QUANTITATIVE PROBLEMS connect

- 12.** You are curious as to whether men tend to like your product more than women do. To learn about this, you collected 22 surveys, asking respondents to rate your product on a scale of one to seven. (LO1)

INDIVIDUAL NUMBER	SEX	RATING
1	Male	2
2	Male	4
3	Female	2
4	Female	2
5	Female	5
6	Male	1
7	Female	3
8	Female	3
9	Male	7
10	Male	3
11	Male	4
12	Female	4

13	Female	6
14	Male	7
15	Female	3
16	Female	1
17	Male	4
18	Male	5
19	Male	6
20	Male	2
21	Female	4
22	Female	5

Using these data, solve for the regression line for a dichotomous treatment that relates your product rating to a person's sex.

149

13. You've collected data on two variables, Y and X , and you are interested in estimating the simple regression line $Y = b + m \times X$. You have the following summary statistics:

$$s\text{Cov}(X, Y) = -18$$

$$s\text{Var}(X) = 3$$

$$s\text{Var}(Y) = 6$$

$$\text{Mean of } Y = 32$$

$$\text{Mean of } X = 12$$

What is the intercept and slope for your regression line? (LO2)

14. Your firm is interested in learning more about how its salaries relate to its employees' tenure with the firm. It has collected the following data for 25 of its employees.

EMPLOYEE NUMBER	TENURE (YEARS)	SALARY (\$)
1	15	53,408
2	32	77,230

3	14	53,664
4	20	55,647
5	25	60,611
6	14	51,991
7	28	71,071
8	30	69,189
9	28	67,359
10	17	50,978
11	14	56,176
12	6	38,865
13	21	58,176
14	11	52,101
15	14	50,941
16	32	73,964
17	29	67,873
18	33	73,860
19	27	60,519
20	16	48,474
21	26	69,574
22	3	34,594
23	14	52,176
24	9	56,444
25	14	57,806

Plot these data points, and describe using regression how salary relates to firm tenure for this group. (LO2)

product rollout. To do this, it collected surveys gauging respondents' perceived likelihood of purchase (out of 100), along with their age and income levels.

RESPONDENT NUMBER	LIKELIHOOD	AGE (YEARS)	INCOME (\$)
1	13	70	96,345
2	30	51	73,096
3	74	34	74,180
4	74	61	78,325
5	54	64	95,851
6	61	62	119,116
7	43	70	98,425
8	4	27	69,385
9	52	69	80,768
10	46	50	57,102
11	75	24	63,703
12	84	25	62,667
13	66	47	69,375
14	32	63	88,791
15	30	56	73,974
16	96	34	64,727
17	63	56	74,500
18	16	20	43,648
19	90	21	64,475
20	60	42	67,863

Using these data, describe using regression how the likelihood of purchase relates to age and income. (LO6)

Correlation vs. Causality in Regression Analysis

6

LEARNING OBJECTIVES

After completing this chapter, you will be able to:

- LO6.1** Differentiate between correlation and causality in general and in the regression environment.
- LO6.2** Calculate partial and semi-partial correlations.
- LO6.3** Execute inference for correlational regression analysis.
- LO6.4** Execute passive prediction using regression analysis.
- LO6.5** Execute inference for determining functions.
- LO6.6** Execute active prediction using regression analysis.
- LO6.7** Distinguish the relevance of model fit between passive and active prediction.

dataCHALLENGE Where to Park Your Truck—Redux

For our data challenge in [Chapter 5](#), we considered the case in which you just purchased a food truck and have begun selling in a large college town. You have been changing location every few days to get a sense of local demand, and you've decided to collect some data. You collect data each day on your revenues and distance (in miles) from the center of the local university.

Suppose you have daily data covering a time span of nine weeks, and you have estimated the following regression equation using data on Revenue and Distance:

$$\text{Revenue} = 918.32 - 56.18 \times \text{Distance}$$

Tomorrow you are considering two different locations, one two miles from the center of the local university, and another that is one mile farther.

Can you use your estimated regression equation to predict how revenues will differ between these two locations, and if so, how?

152

Introduction

At the end of [Chapter 1](#), we made the distinction between passive and active prediction. As a reminder, prediction is passive when done with no exogenous alteration of variables, as in making a weather prediction following a given weather pattern. Prediction is active when done with exogenous alteration of variables, as in making a profit prediction for a chosen price point. In this chapter, we elaborate on this distinction in general and within the regression framework. We explain how passive prediction is rooted in correlation, and we provide the reasoning that allows one to properly utilize a regression model to make a passive prediction. We then analogously explain how active prediction

is rooted in causation, and we provide the reasoning that allows one to properly utilize a regression model to make an active prediction.

A deep understanding of the difference between correlation and causality is immensely important in business and well beyond. The news is teeming with stories about studies conducted in medicine, education, psychology, and elsewhere that loosely interpret the implications of the analysis that was conducted. By developing and understanding the assumptions requisite for different types of prediction, this chapter will help you to critically assess such studies. You will be able to determine whether they are appropriate for active prediction, passive prediction, or neither. Improper use of data analysis for making predictions in business can have significant consequences for the health and even survival of a firm. This chapter is designed to provide the reasoning tools to ensure such errors do not take place under your watch.

The Difference Between Correlation and Causality

LO 6.1 Differentiate between correlation and causality in general and in the regression environment.

Consider the following scenario. You are the head of the sales division of your firm, and you recently received information from an online professional university, OnlineEd, advertising its sales training program. When emphasizing the benefits of the program, OnlineEd notes that firms utilizing the program had, on average, 15% higher sales the following year than peer firms that did not utilize the program. How should a statistic such as this factor into your assessment of the program's value?

To assess this claim, we should begin by establishing the two variables about which information is being provided. The first is the one-year change in sales for a given firm; call this variable SalesChng. The second is the

training program; call this variable TrainProg. SalesChng can take on a range of numerical values (e.g., 10.2, -8.7, etc.), representing the percentage change in sales for a given year. TrainProg can take on the values zero and one, indicating whether or not the firm participated in the program the previous year.

The information provided by OnlineEd, in terms of our newly defined random variables, is:
 $\overline{\text{SalesChng}|\text{TrainProg} = 1} - \overline{\text{SalesChng}|\text{TrainProg} = 0} = 15$. That is, the difference in

153

the average change in sales for those who used the program and the average change in sales for those who did not use the program is 15%. As a potential client of OnlineEd, we are likely to be interested in whether there is a causal impact of the training program on a firm's subsequent sales. Here, SalesChng is the outcome, TrainProg is the treatment, and we would like to measure the average treatment effect (ATE) of TrainProg on SalesChng, where $\text{ATE} = E[\text{SalesChng}_i^T - \text{SalesChng}_i^{NT}]$, as we discussed in [Chapter 4](#).

That is, we would like to know, for a given firm, what is the expected difference in its sales growth between receiving the treatment (sales training) and not receiving the treatment. We know from [Chapter 4](#) that assuming a random sample of firms from the population and random treatment assignment allows us to use the statistic given to us, $\overline{\text{SalesChng}|\text{TrainProg} = 1} - \overline{\text{SalesChng}|\text{TrainProg} = 0}$, as an unbiased estimate of the ATE.

Thus far, our discussion of causality (from [Chapter 4](#)) has been entirely in the context of a single, dichotomous treatment, for which we measure the ATE. However, as we discussed in [Chapter 5](#), treatments can take on many levels and dimensions. Therefore, as we consider causality more generally in this chapter and beyond, we need a broader framework for modeling it. In fact, the framework is quite straightforward and builds on a topic we briefly introduced in [Chapter 1](#): the data-generating process.

Recall that the data-generating process is defined as the underlying mechanism that produced the pieces of information contained in a dataset. The data-generating process, as defined, expresses a causal relationship among variables. In particular, we write one variable as a function of one or more other variables, implying that its realized values depend on, that is, are caused by, these other variables.

We can express a data-generating process in general terms as follows:

$$Y_i = f_i(X_{1i}, X_{2i}, \dots, X_{Ji})$$

Here, Y plays the role of the outcome, and X_1, \dots, X_J play the roles of treatments. In practice, we seldom explicitly consider, or even observe, all of the treatments that causally determine the outcome. Consequently, it is common practice to separate out the treatments we explicitly consider and those we don't, resulting in a formulation for the data-generating process as follows:

$$Y_i = f_i(X_{1i}, X_{2i}, \dots, X_{Ki}, U_i)$$

Within this formulation, X_1 through X_K is the subset of treatments we are considering (and so $K \leq J$), and U is a conglomeration of all the treatments we are not considering, i.e., it consists of “all other factors affecting Y .” Lastly, it is extremely common to separate out this last “other factors” term as follows:

$$Y_i = f_i(X_{1i}, X_{2i}, \dots, X_{Ki}) + U_i$$

In words, we have that our outcome variable Y is determined by a function of

treatments we are considering (X_1 through X_K), plus the combined effect of all other treatments affecting Y that we are not explicitly considering. We define $f_i(X_{1i}, X_{2i}, \dots, X_{Ki})$ as the **determining function**, since it comprises the part of the outcome that we can explicitly determine (as opposed to U_i , which can only be inferred by solving $Y_i - f_i(X_{1i}, X_{2i}, \dots, X_{Ki})$).

determining function The part of the outcome that we can explicitly determine, $f_i(X_{1i}, X_{2i}, \dots, X_{Ki})$.

154

Let's now apply this data-generating process framework to our OnlineEd example. Here, SalesChng plays the role of Y and TrainProg plays the role of X . So, we have the following data-generating process for the change in sales:

$$\text{SalesChng}_i = f_i(\text{TrainProg}_i) + U_i$$

This formulation indicates that a firm i 's change in sales for a given year is equal to a function of whether it used the sales training program, plus the combined effect of all other factors influencing its change in sales. The causal effect of the training program for a given firm is the difference in its sales growth when TrainProg changes from 0 to 1. Using our framework for the data-generating process, the causal effect of the training program is:

$$f_i(1) + U_i - (f_i(0) + U_i) = f_i(1) - f_i(0)$$

Hence, when working with the data-generating process, the causal effect for a given variable on the outcome boils down to its impact on the determining function.

There are two noteworthy features of our data-generating process as a

framework for modeling causality. First, note that the reasoning we established to measure an average treatment effect using sample means easily maps into this framework. For the OnlineEd example, consider the following:

1. $\text{ATE} = E \left[\overline{\text{SalesChng}_i^T} - \overline{\text{SalesChng}_i^{NT}} \right] = E [f_i(1) + U_i] - E [f_i(0)]$. Hence, the average treatment effect is just the expected change in the determining function (across all individuals) when changing the treatment status.

2. If we assume a random sample, we know $\overline{\text{SalesChng}|\text{TrainProg} = 1}$ is an unbiased estimator for $E[\text{SalesChng}|\text{TrainProg} = 1]$, which equals $E[f_i(1)] + E[U_i|\text{TrainProg} = 1]$ after some simple substitution and algebra.
3. If we assume a random sample, we also know $\overline{\text{SalesChng}|\text{TrainProg} = 0}$ is an unbiased estimator for $E[\text{SalesChng}|\text{TrainProg} = 0]$, which equals $E[f_i(0)] + E[U_i|\text{TrainProg} = 0]$ again after some simple substitution and algebra.
4. Combining 2 and 3 above, we have that $\overline{\text{SalesChng}|\text{TrainProg} = 1} - \overline{\text{SalesChng}|\text{TrainProg} = 0}$ is an unbiased estimator of $E[f_i(1)] + E[U_i|\text{TrainProg} = 1] - E[f_i(0)] - E[U_i|\text{TrainProg} = 0]$.
5. If we assume random treatment assignment, then conditioning on whether the training program was used has no impact. Consequently, $E[U_i|\text{TrainProg}=1] = E[U_i]$ and $E[U_i|\text{TrainProg}=0] = E[U_i]$. Combining this fact with some simple substitution and algebra, we have $E[f_i(1)] + E[U_i|\text{TrainProg}=1] - E[f_i(0)] - E[U_i|\text{TrainProg}=0] = E[f_i(1)] - E[f_i(0)]$.
6. Combining points 1–5, assuming a random sample and random treatment assignment implies that $\overline{\text{SalesChng}|\text{TrainProg} = 1} - \overline{\text{SalesChng}|\text{TrainProg} = 0}$ is an

unbiased estimator of the expected difference in the determining function ($E[f_i(1) - f_i(0)]$), which is the average treatment effect (ATE).

The second noteworthy feature of our data-generating process as a framework for causal analysis is that this framework easily extends into modeling causality for multi-level treatments and multiple treatments. If, for example, TrainProg instead measured the number of sales training courses taken, and so took on values of 0, 1, 2, etc., we no longer can utilize

155

the idea of the average treatment effect to measure the causal effect of the program. In this scenario, we have a multi-level treatment, and must then consider the causal impacts of different dosage levels (number of courses taken), not just the effect of a single dose (whether a course was taken).

However, we can easily model the causal effect of a multi-level treatment with a determining function in a data-generating process. Here, the model looks just as before: $SalesChng_i = f_i(TrainProg_i) + U_i$. The causal effect of the training program is completely captured by $f_i(\cdot)$; if we know this function, we can assess the causal impact on SalesChng for any change in TrainProg. For example, if the number of courses changes from 1 to 2, the causal impact is $f_i(2) - f_i(1)$, and if the number of courses changes from 3 to 5, the causal impact on SalesChng is $f_i(5) - f_i(3)$.

For the case of multiple treatments, we can use the determining function to assess the causal impact of a change in one or more treatments. Using our more general model, suppose we have $Y_i = f_i(X_{1i}, X_{2i}) + U_i$. Again, if we know $f_i(\cdot, \cdot)$, then we can use it to determine the causal effects of changes in one or both of our treatments. For example, if X_1 changes from 2 to 4 and X_2 remains constant at 6, then the causal impact of the change in X_1 is $f_i(4, 6) - f_i(2, 6)$. As another example, if X_1 changes from 11 to 3 and X_2 changes from 2 to 5, the causal impact in these changes on Y is $f_i(11, 2) - f_i(3, 5)$.

Now that we have elaborated on the data-generating process and the

determining function, we can express the idea of causal inference (introduced in [Chapter 1](#)) in a more specific way. Causal inference for a given treatment(s) involves establishing, and often estimating, the determining function within a data-generating process. Data analysis used for the purpose of measuring causality among variables involves estimating a determining function, a process we discuss in detail later in this chapter.

6.1

Demonstration Problem

Consider the following two data-generating processes:

1. $Y_i = 4X_{1i} - 2X_{2i} + U_i$
2. $Y_i = X_{1i}^2 + 3X_{1i} - X_{2i}^3 + U_i$

For each data-generating process, answer the following questions:

- a. What is the determining function?
- b. What is the effect on Y that results from an increase in X_1 from 2 to 5?
- c. What is the effect on Y that results from a decrease in X_2 from 6 to 4?
- d. Derive the formula for the change in Y with respect to a change in X_1 .

(*Hint:* This involves solving for a partial derivative).

Answer:

- a. (1) $f(X_{1i}, X_{2i}) = 4X_{1i} - 2X_{2i}$; (2) $f(X_{1i}, X_{2i}) = X_{1i}^2 + 3X_{1i} - X_{2i}^3$
- b. (1) Y increases by 12 ($5 \times 4 - 2 \times 4 = 12$); (2) Y increases by 30 ($5^2 + 3 \times 5 - (2^2 + 3 \times 2) = 30$)
- c. (1) Y increases by 4 ($-2 \times 4 - (-2) \times 6 = 4$); (2) Y increases by 152 ($-4^3 - (-6^3) = 152$)
- d. (1) 4; (2) $2X_{1i} + 3$

A causal relationship between two variables clearly implies co-movement. That is, if X causally impacts Y , then when X changes, we expect a change in Y . However, variables often move together even when there is no causal relationship between them. As a simple example, we may measure the height of two different children between the ages of 5 and 10. Since both children are growing during these ages, their heights will generally move together; however, this co-movement is not due to causality—an increase in height by one child will not cause a change in height for the other.

We are already familiar with how to measure co-movement between two variables in a dataset; this is captured through their sample covariance or correlation, defined again as follows:

$$s\text{Cov}(X, Y) = \frac{\sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})}{N-1}$$

$$s\text{Corr}(X, Y) = \frac{s\text{Cov}(X, Y)}{\sqrt{s\text{Var}(X) \times s\text{Var}(Y)}}$$

Note that, since the denominator of the sample correlation is always positive, these two measures always have the same sign. Therefore, if we say X and Y are positively correlated, this means they also have positive covariance, and vice versa.

When there are more than two variables, e.g., Y , X_1 , and X_2 , we can also measure what's called the *partial correlation between two of the variables*, which is the correlation between the two variables after controlling for at least one other variable. The partial correlation between two variables is their correlation after holding one or more other variables fixed. The partial correlation between Y and X_1 controlling for X_2 is written as $\text{pCorr}(Y, X_1; X_2)$ and measures the correlation between Y and X_1 if we hold X_2 fixed. We will revisit partial correlation in the next section, and provide a formal description of how it is calculated there.

For any two variables, no matter how seemingly unrelated, their correlation is seldom zero. For example, the correlation between monthly mean global temperature anomalies (the difference between the actual

temperature and the mean for each day) and monthly U.S. population between December 2012 and November 2013 is -0.081 . However, it is difficult to imagine a causal relationship between these variables, and by no means does the non-zero correlation we find imply there is a causal relationship.

In fact, we can both have correlation without causation and even have causation without correlation. To see how the latter could occur, consider the following. For a given city, an increase in tax revenue will cause the city's deficit to decline, and vice versa. However, suppose the city council behaved in such a way that every time tax revenue changed, it adjusted spending in the same direction by the same amount. In this scenario, tax revenue causally impacts the deficit, but there would be zero sample correlation between them.

So, what is the difference between causality and correlation? Fundamentally, *causality implies that a change in one variable or variables causes a change in another, while correlation implies that variables move together*. Another way of making a distinction centers on the data-generating process. Data analysis attempting to measure causality generally involves an attempt to measure the determining function within the data-generating process. Data analysis attempting to measure correlation is not concerned about the data-generating process and determining function, and instead uses standard statistical formulas (sample correlation, partial correlation) to assess how variables move together.

COMMUNICATING DATA 6.1

PHYSICAL FITNESS AND ACADEMIC SUCCESS

A recent study of Kansas elementary and middle school students found that students who were able to attain fitness goals in five areas scored notably higher on mathematics and reading tests compared to students who were not physically fit. What does this finding tell us? First, it tells us that physical fitness

and academic performance are positively correlated. Consequently, if we come across a student who is physically fit, chances are he or she performs relatively well in mathematics and reading.

Does this finding mean physical fitness causally impacts academic performance? If we were able to express the data-generating process for academic performance as: $AP_i = f_i(PF_i, X_{2i}, \dots, X_{Ki}) + U_i$, where AP is academic performance and PF is physical fitness, then we can conclude there is a causal impact of physical fitness on academic performance. However, the findings of this study speak only to correlation. The positive correlation the researchers found does not necessarily imply AP is increasing in PF in the context of the data-generating process.

How could we have a positive correlation between physical fitness and academic performance but have no causal relationship (or even a negative causal relationship) between physical fitness and academic performance? As an example, suppose students who are physically fit tend to come from families that create a regimented environment, where time is set aside both for physical activities and for studying. Suppose also that a regimented studying environment leads to higher academic performance. Then, a positive correlation between physical fitness and academic performance could be due to their mutual relationship with regimented studying, despite there being no causal impact of physical fitness on academic performance.

As we'll see in the next two sections, we can use regression to measure correlations and causality. And depending on its purpose and ability to provide a proper measurement, we can use the estimated regression equations for different types of prediction. We turn to this topic and distinction next.

Regression Analysis for Correlation

LO 6.2 Calculate partial and semi-partial correlations.

In [Chapter 5](#), we introduced regression analysis as a means of establishing a line, or other function, that “best” describes a given data set. In this section, we explain when and how the function used to describe the data sample can be used to describe the population from which the sample was taken. When it is believed that a regression equation “best” describes the population, it is a popular, and can be a highly effective, tool for performing passive prediction, as we describe below. However, a regression equation that best describes the population is not necessarily a good tool for making active predictions, a point on which we elaborate in the next section.

REGRESSION AND SAMPLE CORRELATION

Suppose you are working at Kellogg’s, and the company is launching its newest breakfast cereal, Honey Wheat Crunch. To date, the launch has been limited to a handful of major cities, but Kellogg’s is interested in rolling out the product on a larger scale. To aid in its

158

TABLE 6.1 Honey Wheat Crunch Store Data

STORE	SALES (BOXES)	AVG. PRICE (\$)	AVG. HH SIZE
1	1085	4.56	2.72
2	564	5.42	2.97
3	1495	3.87	3.42
4	592	4.99	1.6
5	989	3.92	2.98
6	544	5.61	3.47
7	837	4.76	1.8
8	1163	4.44	3.26
9	1027	4.32	2.89
10	947	5.14	1.92

...

...

...

...

decision, it has been analyzing grocery store data containing information about its sales, prices, and customer demographics. The dataset is a cross-section of 230 stores, illustrated (in part) in [Table 6.1](#).

Here, Sales is the number of boxes of Honey Wheat Crunch sold in a given month, AvgPrice is the average price for a box of Honey Wheat Crunch for that store during the observed month, and AvgHHSIZE is the average size of the households for customers at that grocery store.

As we discussed in [Chapter 5](#), we can use regression analysis to describe these data. For example, we might want to describe the relationship between Sales and the other two variables, and a standard choice of function to describe this relationship takes the form:

$$\text{Sales} = b + m_1 \text{AvgPrice} + m_2 \text{AvgHHSIZE}$$

After solving the sample moment equations for the full dataset (available through Connect), we get the following equation:

$$\text{Sales} = 1591.54 - 181.66 \times \text{AvgPrice} + 128.09 \times \text{AvgHHSIZE}$$

To this point, we have considered this solution to the sample moment equations as only providing us the equation that best describes the data. However, it tells us more than this. To begin, it provides information about how the variables in the equation are correlated within our sample. Before explaining the specific information the regression provides in this regard, let's briefly consider different ways one can measure correlation between two variables.

The standard measure of correlation is the **unconditional correlation** between two variables X and Y . This is simply a scaled version of the sample

covariance, with the formula: $\text{Corr}(X, Y) = \frac{\text{sCov}(X, Y)}{S_X \times S_Y}$. Here, S_X is the sample standard deviation for X and S_Y is the sample standard deviation for Y . The unconditional correlation is always between -1 and 1 . If it is positive, it implies X and Y generally move in the same direction (if X increases, Y tends to increase); if it is negative, it implies X and Y generally move in opposite directions.

unconditional correlation The standard measure of correlation.

159

Often it is the case that two variables move together because they are both strongly related to a third variable. For example, we may want to know, for flights going from Chicago to Atlanta, the correlation between the departure delays for United (X) and the departure delays for American Airlines (Y). However, we may not be interested in correlation that is due to weather, measured as rainfall in Chicago on the day of the flight (Z). Therefore, we may instead measure a **partial correlation** between X and Y , briefly introduced in the previous section.

partial correlation The partial correlation between X and Y is a measure of the relationship between these two variables, holding at least one other variable fixed.

Informally, the partial correlation between X and Y is a measure of the relationship between these two variables, holding at least one other variable fixed. More formally, the partial correlation between X and Y , controlling for Z , is the unconditional correlation between the residuals from regressing X on Z and from regressing Y on Z . We have $\text{pCorr}(X, Y; Z) = \text{Corr}(e^{X,Z}, e^{Y,Z})$, where $e^{X,Z}$ are the residuals when regressing X on Z and $e^{Y,Z}$ are the residuals when regressing Y on Z .

For our airline example, suppose regressing X on Z yields: $X = 2.4 + 1.6 \times Z$. And, regressing Y on Z yields $1.9 + 2.1 \times Z$. Table 6.2 contains eight observations of delays for United and American, along with the weather

(rainfall). The residuals are simply the difference between the observed delay and the corresponding point on the regression line for that airline (e.g., the first United residual is $5 - (2.4 + 1.6 \times (0.5)) = 1.8$). We then calculate the unconditional correlation between these residuals, and this is the correlation between X and Y , controlling for Z .

A measure of correlation that is closely related to partial correlation is **semi-partial correlation**. While partial correlation holds another variable(s), Z , fixed for both X and Y , sometimes we want to hold Z fixed for only X or only Y . When we do this, we are measuring semi-partial correlation between X and Y , which we write as $\text{spCorr}(Y, X(Z))$ when we hold Z constant just for X . To calculate semi-partial correlation between Y and X , holding Z constant for X , we calculate the unconditional correlation between Y and the residuals from regressing X on Z . This means we would calculate the unconditional correlation between United delays (the first column of [Table 6.2](#)) and the American residuals (the last column of [Table 6.2](#)).

semi-partial correlation The semi-partial correlation between X and Y is a measure of the relationship between these two variables, holding at least one other variable fixed for only X or Y .

How do the correlations detailed above relate to regression analysis? Very closely, it turns out. For the general regression equation $Y = b + m_1X_1 + \dots + m_KX_K$, the solutions

TABLE 6.2 Airline Delays and Regression Residuals

UNITED DELAY (MINS.)	AMERICAN DELAY (MINS.)	WEATHER (INCHES OF RAIN)	UNITED RESIDUAL	AMERICAN RESIDUAL
5	0	0.5	1.8	-2.95
25	19	1.8	19.72	13.32
0	8	0	-2.4	6.1
2	0	0.3	-0.88	-2.53

14	21	1.5	9.2	15.95
0	7	0.4	-3.04	4.26
3	6	0	0.6	4.1
11	8	1.1	6.84	3.79

160

we get for m_1 through m_K when solving the sample moment equations are proportional to the partial and semi-partial correlation between Y and the respective X s. For example, the solution for m_1 is: $m_1 = c \times \text{spCorr}(Y, X_1 | X_2, \dots, X_K)$, where c is a positive constant. Consequently, regression analysis provides direct information about partial correlations, and immediately allows us to determine whether they are positive or negative. Returning to our Honey Wheat Crunch example, our regression equation of Sales = 1591.54 - 181.66 × AvgPrice + 128.09 × AvgHHSIZE implies that the (semi-)partial correlation between Sales and AvgAge is negative and the (semi-)partial correlation between Sales and AvgHHSIZE is positive.

6.2

Demonstration Problem

For the data in [Table 6.3](#), solve for:

a. The partial correlation between:

- i. Y and X_1 , controlling for X_2 and X_3 ($\text{pCorr}(Y, X_1; X_2, X_3)$)
- ii. Y and X_2 , controlling for X_1 and X_3 ($\text{pCorr}(Y, X_2; X_1, X_3)$)
- iii. Y and X_3 , controlling for X_1 and X_2 ($\text{pCorr}(Y, X_3; X_1, X_2)$)

b. The semi-partial correlation between:

- i. Y and X_1 , controlling for X_2 and X_3 ($\text{spCorr}(Y, X_1 | X_2, X_3)$)
- ii. Y and X_2 , controlling for X_1 and X_3 ($\text{spCorr}(Y, X_2 | X_1, X_3)$)
- iii. Y and X_3 , controlling for X_1 and X_2 ($\text{spCorr}(Y, X_3 | X_1, X_2)$)

c. The regression equation: $Y = b + m_1X_1 + m_2X_2 + m_3X_3$

TABLE 6.3 Data for Y , X_1 , X_2 , and X_3

Y	X_1	X_2	X_3
507	10	44	83
151	2	9	20
412	8	35	67
284	4	19	41
411	8	36	67
537	10	48	87
455	8	36	71
421	8	36	68
215	2	13	30
147	3	8	18
364	6	31	59
638	12	53	102
375	5	29	58
275	6	29	49
604	12	56	102

505	10	43	81
404	6	34	65
408	8	33	63
233	4	19	36
449	11	44	77
395	9	37	64
217	2	14	30
263	4	20	39
232	3	16	32

Answer:

a. Partial correlations:

- i. Regress Y on X_2 and X_3 and collect the residuals.
Regress X_1 on X_2 and X_3 and collect the residuals. The correlation between these residuals is 0.279.
- ii. Regress Y on X_1 and X_3 and collect the residuals.
Regress X_2 on X_1 and X_3 and collect the residuals. The correlation between these residuals is -0.787.
- iii. Regress Y on X_1 and X_2 and collect the residuals.
Regress X_3 on X_1 and X_2 and collect the residuals. The correlation between these residuals is 0.956.

b. Semi-partial correlations:

- i. Regress X_1 on X_2 and X_3 and collect the residuals. The correlation between Y and the residuals is 0.017.
- ii. Regress X_2 on X_1 and X_3 and collect the residuals. The correlation between Y and the residuals is -0.074.

- iii. Regress X_3 on X_1 and X_2 and collect the residuals. The correlation between Y and the residuals is 0.187.
- c. The estimated regression equation is: $Y = 24.12 + 3.25X_1 - 8.41X_2 + 9.97X_3$

Note that the partial correlations and the regression coefficients all have the same sign for each combination of Y and the X 's.

REGRESSION AND POPULATION CORRELATION

LO 6-3 Execute inference for correlational regression analysis.

Besides providing information on (partial) correlation in our sample, the regression equation we've solved for our sample can also provide information about a broader population. Consider again our earlier Honey Wheat Crunch example. As we noted, the sample contains information for 230 different stores at a given point in time (a single month). This sample comes from a population that we might care about. We might want to learn about the population in a way that allows us to make meaningful predictions.

Before we can start making predictions for members of our population, we must first define the population. We may define the population as the set of all grocery stores in

162

the United States during that month, or the set of all grocery store/month combinations over a two-year span. As we will see below, how we define the population from which our sample was drawn will impact how and when we can apply our analysis of the sample (e.g., regression results).

Suppose we define the population as all grocery stores in the United States during the month of our sample. Now, imagine we had data for this entire population; that is, we observed the Sales, AvgPrice and AvgHHSIZE for each grocery store in the United States during this single month. Then,

just as in our sample, we could use a regression equation to describe the entire population.

$$\text{Sales} = B + M_1 \text{AvgPrice} + M_2 \text{AvgHHSIZE}$$

Here, we use capital letters to indicate that these are the intercept and slopes for the population, rather than a sample.

If we had data for the entire population, we could simply solve for B , M_1 , and M_2 by solving the sample moment equations using the entire population of data. Our solution would yield the equation that best describes the entire population. Of course, we generally do not have the entire population of data; we only have a sample. In [Chapter 5](#), we used regression analysis as a tool to describe the sample. However, we can also use regression analysis on our sample to learn about the corresponding regression equation for the entire population.

The challenge we face is a specific case of our discussion on population parameters and estimators in [Chapter 3](#). Here, the population parameters are B , M_1 , and M_2 ; they are the intercept and slopes for the regression equation that best describes the population, and we cannot directly solve for them since we do not have access to the entire population of data. However, we do have a sample of data, and we can solve for the intercept and slopes for the regression equation that best describes our sample. We write the regression equation for the sample as $\text{Sales} = b + m_1 \text{AvgAge} + m_2 \text{AvgHHSIZE}$, and solve for b , m_1 , and m_2 . Just as the sample mean is an estimator for the population mean for a random variable, the intercept and slope(s) of the regression equation describing a sample are estimators for the intercept and slope(s) of the corresponding regression equation describing the population. For our breakfast cereal example, b , m_1 , and m_2 are the estimators and B , M_1 , and M_2 are the population parameters they are used to estimate, respectively.

Let's now generalize this idea. Suppose we have a population of data consisting of Y, X_1, X_2, \dots, X_K . Suppose also that (B, M_1, \dots, M_K) are the population parameters such that $Y = B + M_1X_1 + \dots + M_KX_K$ best describes the population among all regression equations of this form (i.e., linear combinations of X_1 through X_K). This implies that (B, M_1, \dots, M_K) solve

$$\frac{\sum_{i=1}^{\text{Pop}} e_i}{\text{Pop}} = \frac{\sum_{i=1}^{\text{Pop}} (Y_i - B - M_1 \times X_{1i} - \dots - M_K \times X_{Ki})}{\text{Pop}} = 0$$

$$\frac{\sum_{i=1}^{\text{Pop}} e_i \times X_{1i}}{\text{Pop}} = \frac{\sum_{i=1}^{\text{Pop}} (Y_i - B - M_1 \times X_{1i} - \dots - M_K \times X_{Ki}) \times X_{1i}}{\text{Pop}} = 0$$

$$\frac{\sum_{i=1}^{\text{Pop}} e_i \times X_{Ki}}{\text{Pop}} = \frac{\sum_{i=1}^{\text{Pop}} (Y_i - B - M_1 \times X_{1i} - \dots - M_K \times X_{Ki}) \times X_{Ki}}{\text{Pop}} = 0$$

where Pop is the size of the entire population.

163

Suppose now that we have a sample of size N from this population. Then, we can analogously solve for the values (b, m_1, \dots, m_K) such that $Y = b + m_1X_1 + \dots + m_KX_K$ best describes this data sample among all regression equations of this form. Hence, (b, m_1, \dots, m_K) solve

$$\frac{\sum_{i=1}^N e_i}{N} = \frac{\sum_{i=1}^N (Y_i - b - m_1 \times X_{1i} - \dots - m_K \times X_{Ki})}{N} = 0$$

$$\frac{\sum_{i=1}^N e_i \times X_{1i}}{N} = \frac{\sum_{i=1}^N (Y_i - b - m_1 \times X_{1i} - \dots - m_K \times X_{Ki}) \times X_{1i}}{N} = 0$$

$$\frac{\sum_{i=1}^N e_i \times X_{Ki}}{N} = \frac{\sum_{i=1}^N (Y_i - b - m_1 \times X_{1i} - \dots - m_K \times X_{Ki}) \times X_{Ki}}{N} = 0$$

Then, (b, m_1, \dots, m_K) are estimators for (B, M_1, \dots, M_K) . Put another way, the regression equation we estimate that best describes our sample is an estimator for the regression equation that best describes the population.

While the intercept and slope(s) that best describe our sample may serve as estimators for the intercept and slope(s) that best describe the population,

are they “reliable” estimators? Would they be “reasonable” guesses for their corresponding population parameters? To answer this question, we first note that the regression intercept and slope(s) (b, m_1, \dots, m_K) for a given sample are random variables. Their values depend on the values of Y and the X s in the sample, and since these change each time we take a different sample, so will the corresponding intercept and slope(s) we estimate. Recognizing (b, m_1, \dots, m_K) as random variables, one way to characterize them as reasonable guesses for their corresponding population parameters is for them to be unbiased. That is, $E(b) = B$, $E(m_1) = M_1$, etc. This is the criterion we used in [Chapter 3](#) to argue that the sample mean is a reasonable guess for the population mean of a random variable.

A similar feature to being unbiased is to be consistent. A random variable is a **consistent estimator** of a population parameter if its realized value tends to get very close to the population parameter as the sample size gets large. For example, our estimator for the intercept, b , is a consistent estimator for the population intercept, B , if the absolute value of the difference between b and B ($|b-B|$) tends to get very small as the sample size (N) gets big. If this is the case, we write this property as $b \rightarrow B$ (b “approaches” B). As we summarize in [Reasoning Box 6.1](#), if we have a random sample, our estimators will in fact be consistent.

consistent estimator An estimator whose realized value gets close to its corresponding population parameter as the sample size gets large.

From [Reasoning Box 6.1](#), we know that our sample being random allows us to make a good guess about the population regression equation using our estimated sample regression equation, particularly when the sample size is large. We illustrate this idea in [Figures 6.1a–6.1c](#). [Figure 6.1a](#) plots a full population of data for two variables, Y and X , where the population size is 200. The blue line is the population regression equation for these data; here $B = 15.78$ and $M = 0.86$. The graphs in [Figure 6.1b](#) plot three random samples of size 10 from the population. The associated regression lines are in orange,

and the blue line is a recreation of the population regression line. The graphs in [Figure 6.1c](#) plot three random samples of size 30 from the population. The associated regression lines are in green, and again the blue line is a recreation of the population regression line. Notice that the green lines more closely resemble the blue (population) line compared to the orange lines. This is a simple illustration of consistency for the regression estimators—as the random

164

REASONING BOX 6.1

CONSISTENCY OF REGRESSION ESTIMATORS FOR POPULATION CORRELATIONS

For a population of all possible realizations of Y, X_1, \dots, X_K , let (B, M_1, \dots, M_K) be the population parameters such that $Y = B + M_1X_1 + \dots + M_KX_K$ best describes the population among all regression equations of this form, and so solve the sample moment equations using the entire population. Let $\{Y_i, X_{1i}, \dots, X_{Ki}\}_{i=1}^N$ be a sample of size N from this population, and let (b, m_1, \dots, m_K) be such that $Y = b + m_1X_1 + \dots + m_KX_K$ best describes this data sample among all regression equations of this form (i.e., they solve the sample moment equations). If $\{Y_i, X_{1i}, \dots, X_{Ki}\}_{i=1}^N$ is a random sample, then (b, m_1, \dots, m_K) are consistent estimators of their corresponding population parameters, (B, M_1, \dots, M_K) . We write this result as:

$$b \rightarrow B$$

$$m_1 \rightarrow M_1$$

...

$$m_K \rightarrow M_K$$

sample gets larger, the estimated regression line more closely resembles the population regression line.

Returning to our breakfast cereal example, we have our estimated regression equation as $\text{Sales} = 1591.54 - 181.66 \times \text{AvgPrice} + 128.09 \times \text{AvgHHSIZE}$. This implies that if we move along the dimension of AvgPrice (hold AvgHHSIZE fixed), the slope of Sales with respect to AvgPrice is -181.66 . If we hold AvgHHSIZE fixed, the points along our estimated regression plane have Sales decreasing 181.66 for every unit increase in AvgPrice. Similarly, if we hold AvgPrice fixed, the points along our estimated regression plane have Sales increasing 128.09 for every unit increase in AvgHHSIZE. We know these are good guesses for their corresponding population parameters from [Reasoning Box 6.1](#). However, we often want to do more than just make a good guess; we often want to test hypotheses and/or build confidence intervals for the population parameters.

FIGURE 6.1A Regression Line for Full Population

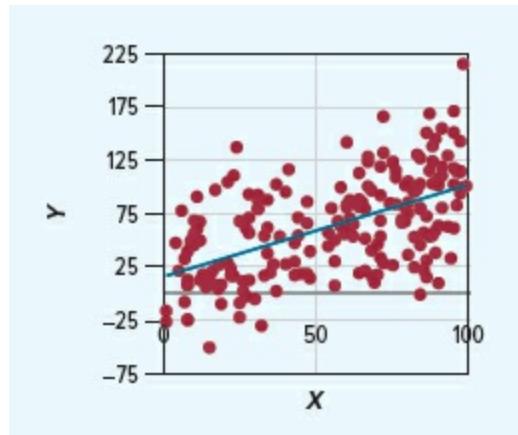


FIGURE 6.1B Regression Lines for Three Samples of Size 10

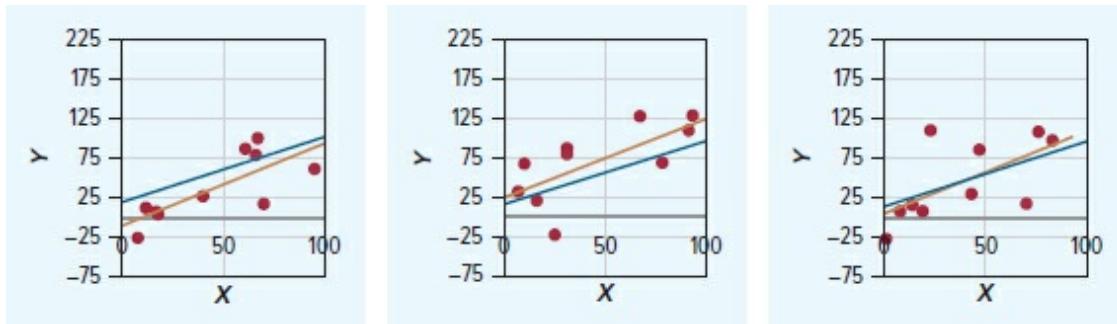
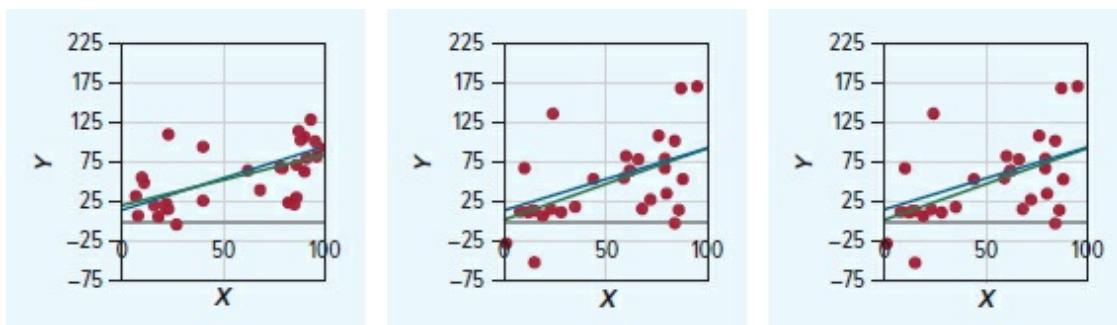


FIGURE 6.1C Regression Lines for Three Samples of Size 30



Just as it was with a population mean in [Chapter 3](#), in order to conduct hypothesis tests or build confidence intervals for the population parameters of a regression equation, we need to know the distribution of our estimators. From [Reasoning Box 6.1](#), we know each estimator becomes very close to its corresponding population parameter for a large sample. Further, just as was the case for our sample mean in [Chapter 3](#), when we have a random sample, the central limit theorem also applies to our regression estimators. Consequently, for a large sample, these estimators are normally distributed. Recall that “large” implied more than 30 observations when we were analyzing the sample mean; however, since we have $K + 1$ estimators when conducting regression analysis, “large” implies at least 30 observations per estimator. Therefore, a “large” sample for regression analysis has at least $30 \times (K + 1)$ observations.

Thus far, we have that a large, random sample implies that:

$$b \sim N(B, \sigma_b)$$

$$m_1 \sim N(M_1, \sigma_{m_1})$$

...

$$m_K \sim N(M_K, \sigma_{m_K})$$

Just as the solution to the sample moment equations provides us with the formulas for (b, m_1, \dots, m_K) for a given sample, it also provides us with the formulas for the standard

166

deviations of these random variables $(\sigma_b, \sigma_{m_1}, \dots, \sigma_{m_K})$. These formulas are not particularly instructive, and are almost always solved via computer, so we do not present them. However, it is useful to note that they generally depend on the X s and the conditional variance of Y given X , i.e., $\text{Var}(Y | X)$. This latter component simply captures how much Y varies around its corresponding point on the regression equation. If we write each element in the population as $Y_i = B + M_1X_{1i} + \dots + M_KX_{Ki} + E_i$, where E_i is the residual, then $\text{Var}(Y | X)$ is simply equal to $\text{Var}(E | X)$. To simplify the analysis, it is a common assumption that this variance is constant across all values of X ; that is, $\text{Var}(Y | X) = \text{Var}(E | X) = \text{Var}(E) = \sigma^2$. This constancy of variance is called **homoscedasticity**.

homoscedasticity Variance constant across all values of X .

In practice, we observe the X s but we do not observe $\text{Var}(E)$. However, we can estimate $\text{Var}(E)$ by calculating the sample variance for the residuals in our sample; call this $\text{Var}(e)$. $\text{Var}(e)$ serves as an estimator for $\text{Var}(E)$ in the same way that S served as an estimator for σ when working with sample means in [Chapter 3](#). Making this substitution, we have $S_b, S_{m_1}, \dots, S_{m_K}$ (calculated from the data using the observed X s and $\text{Var}(e)$) as estimators of $\sigma_b, \sigma_{m_1}, \dots, \sigma_{m_K}$, respectively.

Now that we know the (normal) distribution of our estimators and can estimate their standard deviations using our data sample, we can build confidence intervals and conduct hypothesis tests for their corresponding

population parameters, in exactly the same way we did for the population mean of a random variable in [Chapter 3](#). To see this, we recreate here Reasoning Boxes 3.2 and 3.4, and apply them to regression analysis for population correlations. This results in Reasoning Boxes 6.2 and 6.3, respectively.

REASONING BOX 6.2

CONFIDENCE INTERVALS FOR CORRELATIONAL REGRESSION ANALYSIS

For a population of all possible realizations of Y, X_1, \dots, X_K , let (B, M_1, \dots, M_K) be the population parameters such that $Y = B + M_1X_1 + \dots + M_KX_K$ best describes the population among all linear regression equations using X_1, \dots, X_K , and so solve the sample moment equations using the entire population. Let $\{Y_i, X_{1i}, \dots, X_{Ki}\}_{i=1}^N$ be a sample of size N from this population, and let (b, m_1, \dots, m_K) be such that $Y = b + m_1X_1 + \dots + m_KX_K$ best describes this data sample among all linear regression equations using X_1, \dots, X_K (i.e., they solve the sample moment equations).

Deductive reasoning:

IF:

1. $\{Y_i, X_{1i}, \dots, X_{Ki}\}_{i=1}^N$ is a random sample
2. The size of the sample is at least $30 \times (K + 1)$
3. $\text{Var}(Y | X) = \sigma^2$

THEN:

The interval consisting of b plus or minus 1.65 (1.96, 2.58) times S_b will contain B approximately 90% (95%, 99%) of the time. The same holds true for m_1, \dots, m_K .

Inductive reasoning:

Based on the observation of b , S_b , and N , B is contained in the interval $(b \pm 1.65(S_b))$. The objective degree of support for this inductive argument is 90%. If we instead use the intervals $(b \pm 1.96(S_b))$ and $(b \pm 2.58(S_b))$, the objective degree of support becomes 95% and 99%, respectively. The same holds true for m_1, \dots, m_K .

REASONING BOX 6.3

HYPOTHESIS TESTING FOR CORRELATIONAL REGRESSION ANALYSIS

For a population of all possible realizations of Y, X_1, \dots, X_K , let (B, M_1, \dots, M_K) be the population parameters such that $Y = B + M_1X_1 + \dots + M_KX_K$ best describes the population among all linear regression equations using X_1, \dots, X_K , and so solve the sample moment equations using the entire population. Let $\{Y_i, X_{1i}, \dots, X_{Ki}\}_{i=1}^N$ be such that $Y = b + m_1X_1 + \dots + m_KX_K$ best describes this data sample among all linear regression equations using X_1, \dots, X_K (i.e., they solve the sample moment equations).

Deductive reasoning:

IF:

1. $\{Y_i, X_{1i}, \dots, X_{Ki}\}_{i=1}^N$ is a random sample
2. The size of the sample is at least $30 \times (K + 1)$
3. $\text{Var}(Y | X) = \sigma^2$
4. $B = c_0$

THEN:

We have $b \sim N(c_0, \sigma_b)$, and b will fall within 1.65 (1.96, 2.58) standard deviations of c_0 approximately 90% (95%, 99%) of the time. This also means that b will differ by more than 1.65 (1.96, 2.58) standard deviations from c_0 (in absolute value) approximately 10% (5%, 1%) of the time.

The same holds true for each of m_1, \dots, m_K when assuming, e.g., $M_j = c_j$.

Inductive reasoning:

Using t-stats. If the absolute value of the t-stat for b ($= \left| \frac{b - c_0}{s_b} \right|$) is greater than 1.65 (1.96, 2.58), reject the deduced (above) distribution for b . Otherwise, fail to reject. The objective degree of support for this inductive argument is 90% (95%, 99%).

The same holds true for m_1, \dots, m_K when testing, e.g., $M_j = c_j$.

Using p-values. If the p -value of the t-stat for b is less than 0.10 (0.05, 0.01), reject the deduced (above) distribution for b . Otherwise, fail to reject. The objective degree of support for this inductive argument is 90% (95%, 99%).

The same holds true for m_1, \dots, m_K when testing, e.g., $M_j = c_j$.

Transposition:

If inductive reasoning leads to a rejection of the distribution for b , reject at least one of the assumptions (1, 2, 3, or 4 above) leading to that distribution. If the sample is large, and there is confidence in a random sample and homoscedasticity, this means rejection of the null hypothesis.

We illustrate how to apply this reasoning in practice in [Demonstration Problem 6.3](#).

6.3

Demonstration Problem

Using data on Y , X_1 , and X_2 in the file **Demo6-34.xlsx**, available at www.mhhe.com/prince1e or via Connect, you estimate the following regression equation:

$$Y = b + m_1X_1 + m_2X_2$$

The results are presented in [Table 6.4](#). Using the regression results in [Table 6.4](#), answer the following questions:

1. Test the hypothesis that the intercept is equal to zero, using a confidence level of 90%. Be sure to provide the reasoning behind your result.
2. Build a 95% confidence interval for M_1 —the population regression coefficient for X_1 . Be sure to provide the reasoning behind your result.

TABLE 6.4 Regression Output for Regression of Y on X_1 and X_2

SUMMARY OUTPUT				
<i>Regression Statistics</i>				
Multiple R	0.801606194			
R Square	0.64257249			
Adjusted R Square	0.63563215			
Standard Error	35.49731246			
Observations	106			
ANOVA				
	df	SS	MS	F
Regression	2	233325.5636	116662.7818	92.58519
Residual	103	129786.0967	1260.059192	
Total	105	363111.6604		
	Coefficients	Standard Error	t Stat	P-value
Intercept	-5.929551533	6.862006432	-0.864113374	0.38953

X1	-1.575541513	0.51167844	-3.079163373	0.00266
X2	5.808291356	0.451934097	12.85207598	3.932461

Answer:

1. We assume that: we have a random sample, the sample size is at least $(2 + 1) \times 30 = 90$, there is homoscedasticity. Note that our second assumption is immediately verified since $N = 106$. Given these assumptions, using a 90% confidence level, we fail to reject the hypothesis that the intercept is equal to zero, since the p -value is above 0.10 (it is 0.3895).
2. We assume that: we have a random sample, the sample size is at least $(2 + 1) \times 30 = 90$, there is homoscedasticity. Note that our second assumption is immediately verified since $N = 106$. Given these assumptions, we are 95% confident that M_1 is between -2.59 and -0.56.

PASSIVE PREDICTION USING REGRESSION

LO 6.4 Execute passive prediction using regression analysis.

Now that we know how to get reliable estimates for how variables move together in the population, we can use this information to make predictions. Consider again a population consisting of information on variables: Y, X_1, \dots, X_K . As we've done previously, let (B, M_1, \dots, M_K) be the population parameters such that $Y = B + M_1X_1 + \dots + M_KX_K$ best describes the population among all regression equations of this form. Then, for a given set of values for X_1 through X_K (call them x_1, \dots, x_K), we can use our consistent estimate of the population regression equation to predict how Y will move with a change in the X s.

To see how such a prediction works in general, let $\{Y_i, X_{1i}, \dots, X_{Ki}\}_{i=1}^N$ be a random sample of size N from the population (where N is “large”), and let (b, m_1, \dots, m_K) be such that $Y = b + m_1X_1 + \dots + m_KX_K$ best describes this data sample. Then, define Δx_k as the observed change in X_k and $\widehat{\Delta Y}$ as the predicted change in Y , given changes in the X s of $\Delta x_1, \dots, \Delta x_k$. Given these definitions, we have: $\widehat{\Delta Y} = m\Delta x_1 + \dots + m_K\Delta x_k$ as the predicted change in Y when we observe changes in the X s of $\Delta x_1, \dots, \Delta x_k$.

Let’s now see how such predictions work using our breakfast cereal example. Recall our population was all grocery stores in the United States in a given month, and our estimated regression equation was $\text{Sales} = 1591.54 - 181.66 \times \text{AvgPrice} + 128.09 \times \text{AvgHHSize}$. Suppose we are considering two stores (Store A and Store B) outside our sample, and want to predict the difference in their Sales in a given month. Further, we know Store A has an average price that is \$0.50 higher than Store B, and Store A has an average household size that is 0.4 lower than Store B. Using this information, we predict the difference in Sales between the two stores is: $\widehat{\Delta \text{Sales}} = -181.66 \times 0.50 + 128.09 \times (-0.4) = -142$. That is, we predict Store A will have 142 fewer Sales than Store B.

Our previous example involved making sales predictions for out-of-sample stores during the same month. However, prediction using regression often involves looking into the future, e.g., what will sales be next month, or three months from now? The process of using regression to make predictions about future outcomes is exactly the same as before. Suppose next month we observe, for a given store, a change in the average price for Honey Wheat Crunch of \$0.45 and a change in average household size of 0.2. Then, we predict sales for that store next month will differ by $\widehat{\Delta \text{Sales}} = -181.66 \times 0.45 + 128.09 \times 0.2 = -56$. That is, we predict sales for that store will be lower by 56.

When making predictions about the future, note that our predictions generally apply to unobserved elements of the population, since we are using consistent estimators of the regression equation that best describes that population. If the population consists of all U.S. stores in a single month, then

it is not appropriate to use our estimates to make predictions about other months; those other months are outside of our population. Consequently, when using correlational regression analysis to make predictions about the future, we must be considering a population that spans across time (e.g., store/months). In addition, we must make another assumption regarding our population regression equation. We must assume the population regression equation that best describes our population also best describes the population for the future time periods we will be predicting. That is, we must assume the partial correlations among the variables in our equation are stable over time. For our breakfast cereal example, suppose the current month is June 2017. If we wish to make predictions about future months (e.g., July–December 2017), we must assume the population regression equation that best describes the population for June 2017 also best describes the population for July–December of 2017. By doing so, we could use our regression equation to make predictions about stores during these future months.

170

We conclude this section by noting that the predictions we described are passive predictions, as introduced in [Chapter 1](#). We take an observed (or hypothetically observed) set of values for the X variables as given, and then use our estimated regression equation to make a prediction about Y . The equation we use to make our prediction simply fits the population well; it is not necessarily the determining function of the data-generating process. Therefore, it is not necessarily appropriate for making predictions when the X s are exogenously altered. For this reason, it is not appropriate to refer to Y as the outcome and the X s as treatments in the context of passive prediction or regression for correlational analysis, as these labels imply a role within a data-generating process where Y has a causal relationship with the X s. Instead, we generically refer to Y as the *dependent variable* and the X s as the *independent variables*. These labels are general, in that they apply to regression equations measuring correlation or causality, while the label of Y as the outcome and X s as treatments is appropriate only for causal analysis.

In the context of our breakfast cereal example, we are making predictions

about sales using information on average price and average household size, treating these two independent variables as outside our control. According to our estimated regression equation ($\text{Sales} = 1591.54 - 181.66 \times \text{AvgPrice} + 128.09 \times \text{AvgHHSIZE}$), if we observe the average price for a store increase by a dollar while average household size remains the same, our prediction for the sales of that store will decrease by 181.66.

It is important to understand that if we are utilizing correlational regression analysis and consequently making a passive prediction, the coefficient on average price is not necessarily the causal effect of price on sales. The figure -181.66 is essentially the partial correlation between Sales and Average Price (the correlation holding average household size constant); this correlation could be due to a causal relationship *and/or* a mutual relationship to one or more other variables. For example, the average age of a store's customers may impact the sales and average price for that store, and thus may be correlated with both of these variables. In such a case, our estimate of 181.66 may be picking up the causal impact of price *and* customer age. If we are passively observing a dollar increase in price, then this change in price likely was accompanied by a change in average customer age as well (since these variables are correlated); therefore, a good prediction will account for the effects of both variables changing, as is the case for our estimate of 181.66.

In contrast, if we take a given store and raise its price by a dollar with all other factors unaltered (i.e., we exogenously alter price by a dollar), we want to know the causal effect of just a change in price; and our measure of 181.66 does not necessarily provide this. In order to get a proper measure of this latter effect, we must take a different approach with our underlying reasoning, which we discuss in the next section.

Regression Analysis for Causality

As noted previously, regressions for correlation can be highly effective tools for performing passive prediction, with many applications. However, such

regressions are not necessarily good tools for making active predictions. To make active predictions, we must perform regression analysis that is suitable for establishing causality. Such regressions rely on a more extensive reasoning process, with additional, carefully considered assumptions. The additional assumptions have conceptual appeal, as they closely resemble the assumptions relied upon in the scientific method from [Chapter 4](#). After fully describing regression

171

analysis for causality, we will illustrate why it is the appropriate approach toward evaluating strategic options in business.

REGRESSION AND CAUSATION

LO 6.5 Execute inference for determining functions.

When we attempt to use regression to measure a causal relationship(s), we can again think of our variables in terms of treatments and outcomes. Establishing the causal impact of a treatment(s) on an outcome generally involves the estimation of the determining function for a data-generating process. This process begins by making an assumption about the structure of the data-generating process. As we touched on earlier in this chapter, it is generally assumed that the data-generating process for an outcome, Y , can be written as $Y_i = f_i(X_{1i}, X_{2i}, \dots, X_{Ki}) + U_i$. In addition, we assume that the determining function can be written as follows:

$$f_i(X_{1i}, X_{2i}, \dots, X_{Ki}) = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_K X_{Ki}$$

We can combine these assumptions into a single assumption; the data-generating process can be written as:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_K X_{Ki} + U_i$$

What does this assumption imply? First, it implies that the determining function can be written as a linear combination of the parameters $\alpha, \beta_1, \dots, \beta_K$. This is a rather benign assumption, which makes the determining function amenable to linear regression analysis (since linear regression is for equations that are linear in the parameters). Note that, as we discussed in [Chapter 5](#), this assumption does not force the determining function to be linear in the treatments. For example, we could have X_1 be Price and then have X_2 be Price²—such a function is clearly not linear in one of the treatments, Price.

Our assumed form for the data-generating process also implies that the causal effect of each treatment is constant across observations; this is captured by the fact that none of the parameters has the subscript i , which would indicate variation across observations. This may seem like a strong assumption, but it actually links very closely to our discussion in [Chapter 4](#) on average treatment effects. Specifically, the data we analyze generally will not allow us to measure the causal effect of a treatment on the outcome for a single observation (e.g., individual). However, by observing outcomes and treatments for many observations (and making crucial assumptions), we can measure the *average* causal effect of a treatment on the outcome, analogous to the average treatment effect (ATE) we try to estimate for experiments. Consequently, even if we don't necessarily believe our assumption that the causal effects of our treatments are exactly the same across observations, it will still be the case that the average causal effect is the same across observations, and this is all the data would allow us to measure anyway (absent further assumptions).

It is important to emphasize that, for this assumed form for the data-generating process, the final term, U_i , is a conglomerate of all other factors, besides X_1 through X_K , that determine the outcome, Y . For example, we may express U_i as:

$$U_i = \beta_{K+1} X_{K+1} + \dots + \beta_J X_J$$

It need not take this exact form, but envisioning it this way will simplify our discussion of several topics later in the book. If we view this term as representing “unobserved” factors,

172

often called the **error term**, in this manner, it helps to illustrate how we can write the data-generating process in multiple ways. For example, we could write the data-generating process as

$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_K X_{Ki} + U_i$, or we could write it as $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_K X_{Ki} + \beta_{K+1} X_{K+1} + V_i$, where $V_i = \beta_{K+2} X_{K+2} + \beta_j X_j$. The way we ultimately express the data-generating process when conducting our regression analysis for causality will depend on the variables we observe, the causal effects we want to estimate, and whether we believe it satisfies important assumptions needed to establish causality. We will detail this process later in this section, and revisit this issue on several occasions in subsequent chapters.

error term Represents “unobserved” factors that determine the outcome.

With our assumed form for the data-generating process, it may seem that it is no different than the population regression equation. Comparing the two, for independent variables X_1 through X_K , we write the observations as:

$$Y_i = B + M_i X_{1i} + \dots + M_K X_{Ki} + E_i \text{ (Correlation Model)}$$

$$Y_i = \alpha + \beta_i X_{1i} + \dots + \beta_K X_{Ki} + U_i \text{ (Causality Model)}$$

These two equations look exactly the same; we simply have different names for the parameters and the last term (residual vs. error term). However, there

is an important difference between the Correlation Model (based on the population regression equation) and the Causality Model (based on the data-generating process) because of what each equation represents. The Correlation Model has residuals (E_i) that satisfy the sample moment equations, meaning they have a mean of zero and are uncorrelated with each of the X s. For this model, we simply plot all the data points in the population and write each observation in terms of the equation that best describes those points. In contrast, for the Causality Model, the data-generating process is the process that actually generates the data we observe, and the determining function need not be the equation that best describes the data; that is, it need not be the case that the error term satisfies the sample moment equations. For example, there could be an additional treatment, X_{K+1} , that affects the outcome but is also correlated with one of the other treatments in our determining function, say X_1 . Hence, we can write $U_i = \beta_{K+1}X_{K+1} + V_i$. In such a case, the causal effect of X_1 is still β_1 as we can see from the determining function. However, the outcome (Y) will be (partially) correlated with X_1 both because of X_1 's causal impact on Y , *and* because X_1 is correlated with another variable (X_{K+1}) that also impacts Y . In essence, the partial correlation we estimate between Y and X_1 would capture the causal effects of *both* X_1 and X_{K+1} , meaning the regression equation (based on partial correlations) would not align with the determining function.

A simple example can help illustrate the distinction between the Correlation Model and the Causality Model. Consider the data in [Table 6.5](#), and assume these comprise the entire population of data on Y , X , and U . Here, we have an outcome Y , an observable treatment X and an unobservable treatment U . These data were generated using the data-generating process of $Y_i = 5 + 3.2X_i + U_i$, meaning we have a determining function of $f(X) = 5 + 3.2X$. In [Figure 6.2](#), we plot Y and X along with the determining function (blue line) and the population regression equation (orange line). This simple example shows the essence of the difference between the Correlation Model and Causality Model. The former describes the data best but need not coincide with the causal mechanism generating the data; the latter provides

the causal mechanism but need not describe

173

TABLE 6.5 Data on Outcome (Y), Observed Treatment (X) and Unobserved Treatment (U)

Y	X	U
14.2	4	-3.6
-1.4	2	-12.8
46	10	9
23.8	1	15.6
55	10	18
27.2	4	9.4

the data best. This distinction is crucial when it comes time for prediction, as we discuss further below.

To further illustrate how the data-generating process and population regression equation may differ, we revisit our breakfast cereal example. Suppose, as before, we observe sales, average price, and average household size. And suppose we are interested in the causal effect of price on sales for our product. We express the data-generating process as:

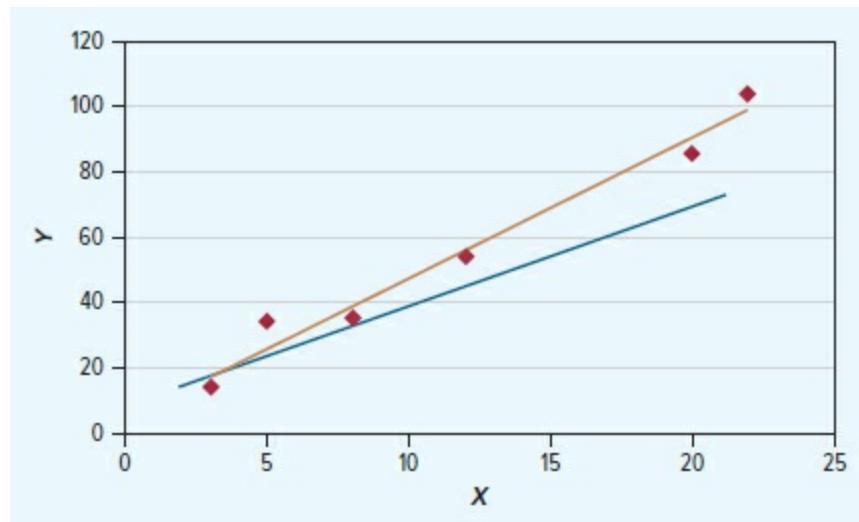
$$\text{Sales}_i = \alpha + \beta_1 \text{AvgPrice}_i + \beta_2 \text{AvgHHSize}_i + U_i$$

Now, suppose the average age of the customer base also impacted sales, and that store managers took the average age of their customers into account when setting their prices. Then, we have $U_i = \beta_3 \text{AvgAge} + V_i$, and AvgPrice and AvgAge would be correlated. In such a scenario, the causal effect of price is β_1 , as expressed through the determining function. However, the partial correlation between sales and average price (holding average

household size fixed) will be a hybrid of the causal effect of average price and the causal effect of average customer age.

To take this further, suppose we observed a store's manager lower average price by one dollar. The causal effect on sales would be for sales to change by $-\beta_1$. However, because price and customer age are correlated, it is likely that the lowered price happened in conjunction with a change in average customer age; say it went down as well, by 1 year. Then,

FIGURE 6.2 Scatterplot, Regression Line, and Determining Function for Y and X



174

REASONING BOX 6.4

EQUIVALENCE OF POPULATION REGRESSION EQUATION AND DETERMINING FUNCTION

IF:

1. The data-generating process for an outcome, Y , can be expressed as:

$$Y_i = \alpha + \beta_1 X_{1i} + \dots + \beta_K X_{Ki} + U_i$$

2. $E[U] = E[U \times X_1] = \dots = E[U \times X_K] = 0$

THEN:

The population parameters (B, M_1, \dots, M_K) , such that $Y = B + M_1 X_1 + \dots + M_K X_K$ best describes the population among all regression equations of this form (i.e., solve the sample moment equations using the entire population) are equal to $(\alpha, \beta_1, \dots, \beta_K)$. In other words, the population regression equation is equal to the determining function of the data-generating process.

the observed change in sales will be the result of the combination of these two changes: $-\beta_1 - \beta_3$. This means sales will be correlated with price in a way that differs from their causal relationship. As we will discuss later in the book, there are ways to assess how partial correlations differ from causal effects, but for now it is sufficient to recognize that they may, and often do, differ.

Now that we have a sense how our population regression equation (measuring correlations) and data-generating process (measuring causation) may differ, we can ask under what circumstances, or assumptions, they are the same. Then, if these assumptions hold, we can simply estimate the population regression equation, and this will provide us the determining function, and thus causal relationship, among our variables. We provide these assumptions in [Reasoning Box 6.4](#).

[Reasoning Box 6.4](#) simply assumes that the data-generating process can be written in the same way as a linear regression equation and that “other factors” besides X_1 through X_K that affect Y have mean zero and are uncorrelated with X_1 through X_K . Then, if we believe these assumptions, the regression equation that has the same form as our determining function, and that best describes the population, is equivalent to that determining function. [Reasoning Box 6.4](#) is highly intuitive. The population regression equation

that best fits the data is defined by the fact that its residuals have mean zero and are uncorrelated with the X s. When assumption #2 holds, this means that this feature mirrors that of the data-generating process—the equation that best describes the data concerning how variables move together (correlations) also best describes the data-generating process (causation).

It is now practical to combine [Reasoning Box 6.1](#) and [Reasoning Box 6.4](#) in order to summarize how we can use a sample to estimate a data-generating process. We do this in [Reasoning Box 6.5](#).

175

REASONING BOX 6.5

CONSISTENCY OF REGRESSION ESTIMATORS FOR DETERMINING FUNCTIONS

For a population of all possible realizations of Y, X_1, \dots, X_K , let $\{Y_i, X_{1i}, \dots, X_{Ki}\}_{i=1}^N$ be a sample of size N from this population. Further, let $(\hat{\alpha}, \hat{\beta}_1, \dots, \hat{\beta}_K)$ be such that $Y = \hat{\alpha} + \hat{\beta}_1 X_1 + \dots + \hat{\beta}_K X_K$ best describes this data sample among all regression equations of this form (i.e., they solve the sample moment equations).

IF:

1. The data-generating process for an outcome, Y , can be expressed as:

$$Y_i = \alpha + \beta_1 X_{1i} + \dots + \beta_K X_{Ki} + U_i$$

2. $\{Y_i, X_{1i}, \dots, X_{Ki}\}_{i=1}^N$ is a random sample
3. $E[U] = E[U \times X_1] = \dots = E[U \times X_K] = 0$

THEN

$(\hat{\alpha}, \hat{\beta}_1, \dots, \hat{\beta}_K)$ are consistent estimators of their corresponding parameters

for the determining function, $Y = \widehat{\alpha} + \widehat{\beta}_1 X_1 + \dots + \widehat{\beta}_K X_K$. We write this result as:

$$\begin{aligned}\widehat{\alpha} &\rightarrow \alpha \\ \widehat{\beta}_1 &\rightarrow \beta_1 \\ &\dots \\ \widehat{\beta}_K &\rightarrow \beta_K\end{aligned}$$

Note that we have made one notational change within [Reasoning Box 6.5](#), in that we now label the estimators from our sample as $(\widehat{\alpha}, \widehat{\beta}_1, \dots, \widehat{\beta}_K)$ rather than (b, m_1, \dots, m_K) . This is to highlight that, when conducting analyses of causality, they are estimating the parameters of a determining function, and not just the parameters of the population regression equation. Consequently, we will use this notation for these estimators for most of what follows, since our focus will be on estimating causal relationships.

If our assumptions establishing the equivalence between the population regression equation and determining function hold, then all our reasoning pertaining to inference for the parameters of the population regression equation will apply to the determining function. Under the proper assumptions, our hypothesis tests and confidence intervals for the parameters of the population regression equation will also apply to the parameters of the determining function. Therefore, we present expanded versions of [Reasoning Box 6.2](#) and [Reasoning Box 6.3](#) in [Reasoning Box 6.6](#) and [Reasoning Box 6.7](#), respectively. These allow us to make inference about causal relationships.

DETERMINING FUNCTION

For a population of all possible realizations of Y, X_1, \dots, X_K , let $\{Y_i, X_{1i}, \dots, X_{Ki}\}_{i=1}^N$ be a sample of size N from this population. Further, let $(\hat{\alpha}, \hat{\beta}_1, \dots, \hat{\beta}_K)$ be such that $Y = \hat{\alpha} + \hat{\beta}_1 X_1 + \dots + \hat{\beta}_K X_K$ best describes this data sample among all linear regression equations using X_1, \dots, X_K (i.e., they solve the sample moment equations).

Deductive reasoning:

IF:

1. The data-generating process for an outcome, Y , can be expressed as:

$$Y_i = \alpha + \beta_1 X_{1i} + \dots + \beta_K X_{Ki} + U_i$$

2. $\{Y_i, X_{1i}, \dots, X_{Ki}\}_{i=1}^N$ is a random sample
3. $E[U] = E[U \times X_1] = \dots = E[U \times X_K] = 0$
4. The size of the sample is at least $30 \times (K + 1)$
5. $\text{Var}(Y | X) = \sigma^2$

THEN:

The interval consisting of $\hat{\alpha}$ plus or minus 1.65 (1.96, 2.58) times $S_{\hat{\alpha}}$ will contain α approximately 90% (95%, 99%) of the time. The same holds true for $\hat{\beta}_1, \dots, \hat{\beta}_K$.

Inductive reasoning: Based on the observation of $\hat{\alpha}$, $S_{\hat{\alpha}}$, and N , α is contained in the interval $(\hat{\alpha} \pm 1.65 (S_{\hat{\alpha}}))$. The objective degree of support for this inductive argument is 90%. If we instead use the intervals $(\hat{\alpha} \pm 1.96 (S_{\hat{\alpha}}))$ and $(\hat{\alpha} \pm 2.58 (S_{\hat{\alpha}}))$, the objective degree of support becomes 95% and 99%, respectively.

The same holds true for $\hat{\beta}_1, \dots, \hat{\beta}_K$

REASONING BOX 6.7

HYPOTHESIS TESTING FOR PARAMETERS OF A DETERMINING FUNCTION

For a population of all possible realizations of Y, X_1, \dots, X_K , let $\{Y_i, X_{1i}, \dots, X_{Ki}\}_{i=1}^N$ be a sample of size N from this population. Further, let $(\hat{\alpha}, \hat{\beta}_1, \dots, \hat{\beta}_K)$ be such that $Y = \hat{\alpha} + \hat{\beta}_1 X_1 + \dots + \hat{\beta}_K X_K$ best describes this data sample among all linear regression equations using X_1, \dots, X_K (i.e., they solve the sample moment equations).

Deductive reasoning:

IF:

1. The data-generating process for an outcome, Y , can be expressed as:

$$Y_i = \alpha + \beta_1 X_{1i} + \dots + \beta_K X_{Ki} + U_i$$

177

2. $\{Y_i, X_{1i}, \dots, X_{Ki}\}_{i=1}^N$ is a random sample
3. $E[U] = E[U \times X_1] = \dots = E[U \times X_K] = 0$
4. The size of the sample is at least $30 \times (K + 1)$
5. $\text{Var}(Y | X) = \sigma^2$
6. $\alpha = c_0$

THEN:

We have $\hat{\alpha} \sim N(c_0, \sigma_\alpha)$ and $\hat{\alpha}$ will fall within 1.65 (1.96, 2.58) standard deviations of c_0 approximately 90% (95%, 99%) of the time. This also means that $\hat{\alpha}$ will differ by more than 1.65 (1.96, 2.58) standard deviations from c_0 (in absolute value) approximately 10% (5%, 1%) of the time.

The same holds true for each of $\hat{\beta}_1, \dots, \hat{\beta}_K$ when assuming, e.g., $\beta_j = c_j$.

Inductive reasoning:

Using t-stats. If the absolute value of the t -stat for $\hat{\alpha}$ ($= \left| \frac{\alpha - c_0}{S_{\hat{\alpha}}} \right|$) is greater than 1.65 (1.96, 2.58), reject the deduced (above) distribution for $\hat{\alpha}$. Otherwise, fail to reject. The objective degree of support for this inductive argument is 90% (95%, 99%).

The same holds true for $\hat{\beta}_1, \dots, \hat{\beta}_K$ when assuming, e.g., $\beta_j = c_j$.

Using p-values. If the p -value of the t -stat for $\hat{\alpha}$ is less than 0.10 (0.05, 0.01), reject the deduced (above) distribution for $\hat{\alpha}$. Otherwise, fail to reject. The objective degree of support for this inductive argument is 90% (95%, 99%).

The same holds true for $\hat{\beta}_1, \dots, \hat{\beta}_K$ when assuming, e.g., $\beta_j = c_j$.

Transposition:

If inductive reasoning leads to a rejection of the distribution for $\hat{\alpha}$, reject at least one of the assumptions (1, 2, 3, 4, 5, or 6 above) leading to that distribution. If there is confidence in assumptions 1–5, this means rejection of the null hypothesis.

We illustrate how to practically apply this reasoning pertaining to causal relationships in [Demonstration Problem 6.4](#).

6.4 Demonstration Problem

Again using data on Y , X_1 , and X_2 in the file **Demo6.34.xlsx**, available at www.mhhe.com/prince1e or via Connect, you estimate the following regression equation:

$$Y = \hat{\alpha} + \hat{\beta}_1 X_1 + \dots + \hat{\beta}_2 X_2$$

The results are presented in [Table 6.5](#) (replicating [Table 6.4](#)). Using the regression results in [Table 6.5](#), answer the following questions:

1. Test the hypothesis that X_1 has no impact on Y (i.e., a change in X_1 will not cause a change in Y), using a confidence level of 99% and a determining function that corresponds to the regression equation above. Be sure to provide the reasoning behind your result.

178

2. Build a 95% confidence interval for the causal impact of X_2 on Y , again using a determining function that corresponds to the regression equation given. Be sure to provide the reasoning behind your result.

TABLE 6.5 Regression Output

SUMMARY OUTPUT				
<i>Regression Statistics</i>				
Multiple R	0.801606194			
R Square	0.64257249			
Adjusted R Square	0.63563215			
Standard Error	35.49731246			
Observations	106			
ANOVA				
	df	SS	MS	F
Regression	2	233325.5636	116662.7818	92.58519
Residual	103	129786.0967	1260.059192	
Total	105	363111.6604		
	Coefficients	Standard Error	t Stat	P-value

Intercept	-5.929551533	6.862006432	-0.864113374	0.38953
X1	-1.575541513	0.51167844	-3.079163373	0.00266
X2	5.808291356	0.451934097	12.85207598	3.93246

Answer:

1. We assume that: the data-generating process for Y can be expressed as:
 $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + U_i$, we have a random sample, $E[U] = E[U \times X_1] = \dots = E[U \times X_K] = 0$, the sample size is at least $(2 + 1) \times 30 = 90$, there is homoscedasticity. Note that our fourth assumption is immediately verified since $N = 106$. Note also that the first and third assumptions are key to ensuring our regression estimates pertain to causality rather than correlation only. Given these assumptions, using a 99% confidence level, we reject the hypothesis that X_1 has no impact on Y , since the p -value is below 0.01 (it is 0.00266).
2. We assume that: the data-generating process for Y can be expressed as:
 $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + U_i$, we have a random sample, $E[U] = E[U \times X_1] = \dots = E[U \times X_K] = 0$, the sample size is at least $(2 + 1) \times 30 = 90$, there is homoscedasticity. Note that our fourth assumption is immediately verified since $N = 106$. Note also that the first and third assumptions are key to ensuring our regression estimates pertain to causality rather than correlation only. Given these assumptions, we are 95% confident that β_2 is between 4.91 and 6.70.

LINKING CAUSAL REGRESSION TO THE EXPERIMENTAL IDEAL

Now that we have established the assumptions that allow us to estimate the parameters of a determining function, we can draw parallels between regression analysis establishing causality and the experimental ideal. In

[Chapter 4](#), we showed that having a random sample and random treatment assignment allowed us to use differences in mean outcomes between the treated and untreated to measure the average treatment effect. Put more simply, a random sample and random treatment assignment allowed us to effectively use our data to measure the causal effect of a treatment. From [Reasoning Box 6.5](#), notice that our use of regression analysis to measure causality relies on two similar assumptions: (1) We have a random sample, and (2) the error terms (U) are uncorrelated with the treatments (Xs). The first assumptions are identical; and the assumption that the errors are uncorrelated with the treatment is analogous to random treatment assignment.

To see the similarity in these two assumptions, recall that random treatment assignment for an experiment implied that the treated group was not “special” compared to the untreated group. This, in essence, means that we wouldn’t expect the group who received the treatment to have different outcomes, on average, if they hadn’t been treated, compared to the untreated group (i.e., no selection bias). It also means that we wouldn’t expect the group who received the treatment to have a systematically different response to the treatment compared to the group who did not receive the treatment (i.e., the effect of the treatment on the treated equals the average treatment effect). While not identical, our assumption of no correlation between the error term (U) and treatment(s) (Xs) is similar to the assumption of random treatment assignment. In essence, this assumption also implies that an observation’s treatment status is not “special,” since it is not related to other factors that influence the outcome (U). Consequently, when we observe differences in the outcome that are correlated with changes in treatment status, we can conclude that these differences are *because* of the changes in treatment status, since other factors affecting the outcome don’t systematically move along with treatment status. We summarize this comparison in [Table 6.6](#), and further illustrate these ideas in [Communicating Data 6.2](#).

ACTIVE PREDICTION USING REGRESSION

LO 6.6 Execute active prediction using regression analysis.

Once we have estimated the determining function for a data-generating process, we can use it to make active predictions. While the determining function can be used to predict a value of Y that corresponds to a given set of values for X , it is most often (and arguably most appropriately) used to predict a *change* in Y that will accompany an exogenous alteration in X . Recall from [Chapter 1](#) that a variable in a dataset is said to be exogenously altered if it changes due to factors outside the data-generating process that are independent of all other variables within the data-generating process.

Within the context of our breakfast cereal example, we may want to know what will be the change in sales when the store manager exogenously decreases price by one dollar. Here, the “factor outside the data-generating process” is the manager’s decision to alter her

TABLE 6.6 Key Assumptions Allowing for Measurement of Causal Relationships

EXPERIMENT	REGRESSION
Random sample	Random sample
Random treatment assignment	Treatment(s) and error term uncorrelated

180

COMMUNICATING DATA 6.2

EXPERIMENTS VS. CAUSAL REGRESSION ANALYSIS

Recently, Facebook was heavily criticized for running an experiment analyzing the effect of the tone of their news feed postings on the tone of users’ statuses. In particular, the company wanted to learn whether showing users news postings that tended to be more positive led to more positive postings and vice versa.

A simplified version of this experiment would involve developing a rating system indicating whether a news feed was positive or not (e.g., $\text{PosFeed} = 1$ if news feed is positive and 0 if negative), and indicating whether an individual's status was positive or not (e.g., $\text{PosStat} = 1$ if status is positive and 0 if negative). For the experiment, we can randomly assign news feeds with different positivity ratings across individuals (treatment) and observe the positivity rating of the individuals' statuses (outcome). Then, we know that if we assume the individuals in the study are a random sample and treatment was randomly assigned, the difference in the positivity in status across the two groups is a reliable measure of the causal effect of news feed positivity.

Consider the same analysis within a regression framework for estimating causality. Here, we assume the data-generating process is $\text{PosStat}_i = \alpha + \beta \text{PosFeed}_i + U_i$. We've already assumed we have a random sample, and random treatment assignment ensures no correlation between unobserved factors affecting status (U) and the news feed (PosFeed). If we further assume that $E[U] = 0$ (easily satisfied when there is a constant term, i.e., α , in the data-generating process), then from [Reasoning Box 6.5](#), we know $\hat{\alpha}$ and $\hat{\beta}$ —the intercept and slope that best describe the sample data—are consistent estimators for the parameters of the determining function. Hence, we see again how running an experiment with random treatment assignment allows us to measure causal relationships.

pricing strategy. Contrast this with the passive predictions we discussed using correlational regression analysis, where we treated the values of our independent variables as being outside our control, and so occurring naturally within the data-generating process.

When we strategically change a single variable affecting an outcome, the effect of that change is the difference in the outcome with and without the change for a given unit of observation. Suppose the determining function is $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + U_i$. Next, suppose we increase X_2 by 1. Then, the change in the outcome caused by this change is:

$$\alpha + \beta_1 X_{1i} + \beta_2 (X_{2i} + 1) + \dots + \beta_K X_{Ki} + U_i - (\alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_K X_{Ki} + U_i) = \beta_2$$

Recall for our breakfast cereal example that our estimated regression equation was $\text{Sales} = 1591.54 - 181.66 \times \text{AvgPrice} + 128.09 \times \text{AvgHHSIZE}$. If we treat this as an estimate of the determining function for Sales, then we would predict that a decrease in average price of one dollar would result in an increase in sales of 181.66.

To make the above prediction, note that we again need to define the population, as we did with passive prediction. Recall that defining the population was important for passive prediction in order to establish the range of heretofore unobserved population elements (grocery stores or grocery-store months) for which we can plausibly make a prediction. In contrast, defining the population is important for active prediction in order to establish the range of population elements for which the data-generating process applies. For active prediction, we may be making predictions concerning elements of the population we already observed, since we are typically trying to predict how the outcome will change

181

with a change in a treatment—something we did not directly observe for any element of the population, even those in the sample.

For our prediction that sales will increase by 181.66 when average price declines by one dollar to be accurate, we must believe that our estimated regression equation is a consistent estimate of the determining function. A key assumption for this to be true is that our independent variables are uncorrelated with the error term. This means that average price must be uncorrelated with other factors that influence sales.

In the previous section, we conjectured that average age of a store's customers may impact sales and the average price for that store. When conducting passive prediction, this relationship among average age, average price, and sales is not consequential—our estimated regression equation still

provides a consistent estimate of partial correlations, which is all we are seeking for passive prediction. However, when conducting active prediction, this relationship is highly consequential. It implies that, within our assumed data-generating process ($\text{Sales} = \alpha + \beta_1 \text{AvgPrice} + \beta_2 \text{AvgHHSIZE} + U$), we have $E[\text{AvgPrice} \times U] \neq 0$. As a result, $\widehat{\beta}_1$ is not a consistent estimator for β_1 , meaning an increase of 181.66 is not a consistent prediction for the change in sales when price exogenously declines by one dollar. This is because $\widehat{\beta}_1$ captures a combination of the causal effect of price *and* the causal effect of average age. Conceptually, stores with a high price are “special” relative to stores with a low price in terms of average age, precluding us from using their differences in sales to determine the causal effect of price on sales.

Of course, we need all of the assumptions in [Reasoning Box 6.5](#) to hold in order for our estimated regression equation to be a consistent estimator for the determining function, and to ultimately engage in active prediction. However, as we discussed in the previous

COMMUNICATING DATA 6.3

WILL DRINKING FATTY MILK MAKE YOU FAT?

It can be tempting to assume that consuming low-fat foods will result in less body fat. However, the analytics on this question tend to suggest otherwise. For example, a study by Swedish experts indicated that, over a 12-year period, middle-aged men who consumed whole milk, cream, and butter had a lower incidence of obesity compared to peers who avoided fattier dairy products. In addition, a European review of 16 studies found the majority to show a lower risk of obesity among people consuming dairy products high in fat. And, on top of this, a recent study showed more weight gain in kids who drank low-fat milk.

Do all these findings imply that consuming fattier dairy products causes a reduction in obesity risk? Perhaps. To be more concrete, we could collect data on individuals, noting whether they were obese ($\text{Obese}_i = 0$ if not obese; $\text{Obese}_i = 1$ if obese) and whether they consumed fattier dairy products

($\text{FatDairy}_i = 0$ if low-fat dairy; $\text{FatDairy}_i = 1$ if fatty dairy). Then, these findings suggest that, if we estimated the regression equation $\text{Obese} = b + m \times \text{FatDairy}$ for these data, we would get a negative estimate for m , indicating that obesity and consumption of fatty dairy products are negatively correlated.

However, for this relationship to be causal, we have to consider the data-generating process. We may write the data-generating process as $\text{Obese}_i = \alpha + \beta \text{FatDairy}_i + U_i$. Key to determining whether our estimate of the regression equation represents a causal relationship is correlation between U_i and FatDairy_i . In short, do we think other factors affecting obesity are correlated with a person's consumption of fatty dairy products? If so, we should be wary of drawing any conclusions about causality. If not, then the negative correlation we found with our regression may be causal. We need only check the other assumptions in [Reasoning Box 6.5](#).

182

subsection, it is our assumptions of a random sample and zero correlation between the error term and treatment(s) that are key toward establishing causality. In the next chapter, and subsequent chapters, we will further discuss the consequences when these assumptions do not hold; we will also discuss remedies, particularly for violations of the latter assumption.

The Relevance of Model Fit for Passive and Active Prediction

LO 6.7 Distinguish the relevance of model fit between passive and active prediction.

We conclude this chapter by making one final comparison between correlational regression analysis for passive prediction and causal regression analysis for active prediction. When we estimate a regression equation of the

form $Y = b + m_1X_1 + \dots + m_KX_K$, we find the values of (b, m_1, \dots, m_K) that solve the sample moment equations, and argue that this gives us the equation that best describes the data. However, our solution is best among only equations of this form, i.e., those that are linear combinations of X_1, \dots, X_K . Suppose instead we wanted to estimate a regression equation using a different set of variables, of the form $Y = c + n_1Z_1 + \dots + n_LZ_L$. For this alternative regression equation, suppose we found the values of (c, n_1, \dots, n_L) that solve the sample moment equations. In both cases, we found the best equation of its type, but between them, which describes the data better?

A simple way of measuring the “fit” of a regression equation, i.e., how well it describes the data, is to calculate its R -squared. Calculating a regression’s R -squared requires two intermediate calculations. First, we calculate the **total sum of squares (TSS)**, defined as:

total sum of squares (TSS) The sum of the squared difference between each observation of Y and the average value for Y .

$$\text{TSS} = \sum_{i=1}^N (Y_i - \bar{Y})^2$$

The TSS is the sum of the squared difference between each observation of Y and the average value for Y . In essence, it is a measure of how much the dependent variable varies. Second, we calculate the **sum of squared residuals (SS_{Res})**, defined as:

sum of squared residuals (SS_{Res}) The sum of the squared residuals.

$$\text{SS}_{\text{Res}} = \sum_{i=1}^N e_i^2$$

The sum of squared residuals should look familiar—this is the expression that is being minimized when we solve the sample moment equations, the

equivalent of engaging in ordinary least squares (OLS). Given our definitions for TSS and SS_{Res} , **R-squared** is defined as:

R-squared The fraction of the total variation in Y that can be attributed to variation in the X s.

$$R^2 = 1 - \frac{SS_{\text{Res}}}{TSS}$$

In words, R -squared is the fraction of the total variation in Y that can be attributed to variation in the X 's. In contrast, the second term in the formula, $\frac{SS_{\text{Res}}}{TSS}$, is the fraction of the total variation in Y that is attributable to the residuals, and thus not attributable to the X s.

Defined this way, R -squared gives us a sense of how well the X s we have in our regression equation describe, or fit, the realizations for Y . A high R -squared implies a good fit, meaning the points on the regression equation tend to be close to the actual Y values. For example, if our estimated regression equation is $Y = 10 + 5 \times X$ and the R -squared is 0.95,

183

this implies only 5% of the variation in Y is attributable to the residuals. Consequently, when, say, $X = 2$, the observed value for Y will tend to be something closer to 20 ($=10 + 5 \times 2$), compared to if the R -squared was instead something smaller, such as 0.4.

The relevance of R -squared for passive and active prediction is quite different. For passive prediction, finding a regression equation with a high R -squared is particularly meaningful. Since we are passively observing the realizations of the independent variable(s) and making a prediction about the dependent variable, we want a regression equation that tends to be close to the realized values of the dependent variable, i.e., has a high R -squared. For such an equation, we expect its predictions to be close to reality.

In contrast, for active prediction, R -squared is less meaningful. To see why, suppose again that our estimated regression equation is $Y = 10 + 5 \times X$ and the R -squared is 0.95. However, suppose we also believe that X was

correlated with the error term in the data-generating process. That is, we assume the data-generating process is $Y = \alpha + \beta X + U$, but believe that X and U are correlated. In such a scenario, we would *not* be able to conclude that 5 is the causal impact on Y from a unit increase in X , and so we could not use this figure to make active predictions about how Y will change when X changes. Therefore, despite a very good fit (R -squared = 0.95), this estimated regression equation is not particularly useful toward active prediction. The strong fit in this case is due to the fact that Y and X are strongly correlated, and this strong correlation is due both to X 's causal effect on Y and due to X 's correlation with other factors (U) that also causally impact Y . Consequently, this equation is well suited for predictions that rely on measures of correlation (passive prediction), but not well suited for predictions that rely on measures of causality (active prediction). For this reason, R -squared is not a primary consideration when evaluating active predictions.

RISING TO THE dataCHALLENGE

Where to Park Your Truck—Redux

Let's return again to the Data Challenge posed at the start of the chapter: where to park your food truck in a large college town. Note that measuring the effect of changing the distance of the truck from the university requires the use of active prediction—you are actively changing the value of one of the X s, rather than observing it as it happens. Consequently, predicting the effect of a change in distance requires us to understand its causal relationship with revenues. We have the estimated regression equation for the sample:

$$\text{Revenue} = 918.32 - 56.18 \times \text{Distance}$$

From what we learned in this chapter, we know the estimate of -56.18 only represents the causal effect of distance on revenue when certain assumptions hold. In particular, we need to assume:

1. The data-generating process for Revenue can be expressed as:

$$\text{Revenue}_i = \alpha + \beta_1 \text{Distance}_i + U_i$$

184

2. $\{\text{Revenue}_i, \text{Distance}_i\}_{i=1}^N$ is a random sample
3. $E[U] = E[U \times \text{Distance}] = 0$

If these assumptions hold, -56.18 is a consistent estimate of the causal effect of distance on revenues, and as such, we should expect revenue to decline by \$56.18 when moving the truck one mile further away. It is crucial to note that assumption 3 implies that “other factors” affecting revenue and the truck’s distance from the campus center are not correlated in the data. This may or may not be true, depending on how location was chosen for the data points in the sample. We investigate issues like this more deeply in the subsequent chapters.

SUMMARY

In this chapter, we have contrasted correlation and causation, and their roles in passive and active prediction. In doing so, we have illustrated the distinction in the necessary assumptions that allow us to learn about population-level correlations vs. causal relationships among variables. Of particular importance toward establishing causality is lack of correlation between “other factors” influencing the outcome/dependent variable and the treatments/independent variables whose effects we are attempting to measure. In subsequent chapters, our focus will be on issues surrounding causal regression analysis and active prediction, and we will discuss myriad ways of dealing with situations where this assumption may not hold.

KEY TERMS AND CONCEPTS

consistent estimator

determining function

error term

homoscedasticity

partial correlation

R-squared

semi-partial correlation

sum of squared residuals

total sum of squares

unconditional correlation

CONCEPTUAL QUESTIONS connect

1. The unconditional correlation between Y and X_1 is 0.72, but the semi-partial correlation between Y and X_1 controlling for X_2 is 0.03. What does this imply about the unconditional correlation between: (LO2)
 - a. Y and X_2 ?
 - b. X_1 and X_2 ?
 - c. Given the above information, does the sign of the unconditional correlation between Y and X_2 have any relationship with the sign of the unconditional correlation between X_1 and X_2 ? (If $\text{Corr}(Y, X_2) > 0$, does this tell us that $\text{Corr}(X_1, X_2) > 0$?)
2. What additional assumptions are needed to conclude that the regression estimators are consistent estimates of the parameters of a determining function, beyond those needed to conclude they are consistent estimators of the parameters of a population regression equation? (LO5)

185

-
3. A dataset on an outcome (Y) and treatment (X) is collected via an experiment, where the treatment is randomly assigned. If we write the data-generating process for Y as: $Y_i = \alpha + \beta X_i + U_i$, what can we say about the correlation between X and U ? (LO1)
 4. Suppose you estimate the following regression equation: $Y = 14 - 3X_1 + 6X_2$. You are willing to assume that you have a random sample and the

sample size is at least 90. Given these assumptions and these estimates, would you predict that Y will decrease by 3 when X_1 is exogenously increased by 1? Why or why not? (LO1)

5. Suppose you estimate the following regression equation for your firm's Sales and Advertising Expenditures for a given month across many regions: $\text{Sales} = 87,142 + 0.12\text{AdExp}$. You are willing to assume that you have a random sample (from a population of region-months spanning the past year into the subsequent year), and the sample size is at least 60. Given these assumptions and these estimates: (LO4)
 - a. Predict the Sales next month if a region is observed to have \$100,000 in advertising expenditure.
 - b. If Region A is observed to spend \$50,000 more than Region B in advertising expenditure, predict the difference in their sales.
 - c. How would your answer to Part b change if we instead asked for a prediction for the increase in Region B's sales if that region increased its advertising expenditure by \$50,000?
6. Regarding R -squared: (LO7)
 - a. What does it measure?
 - b. Why is its magnitude of little relevance when estimating determining functions?
7. Suppose you have a large, random sample of the variables Y and X . You then regress Y on X and get the following results (with standard errors in parentheses): (LO3)

$$Y = 15.2 - 3.7X$$

$$(4.3) (1.2)$$

- a. The numbers 15.2 and -3.7 are the realized values for the intercept and slope (respectively) of the regression equation describing the sample, which are consistent estimators for what population parameters?

- b. Provide a 99% confidence interval for each estimator's corresponding population parameter.
8. Refer to Question 7. (LO4)
- Provide an example of a passive prediction using the above regression results.
 - Provide an example of an active prediction using the above regression results.
 - Is this regression suitable for passive prediction, active prediction, both, or neither?
9. Suppose you have regressed your firm's weekly sales on the number of weekly television ads you've placed for your product. The regression results are as follows: (LO6)
-

$$\text{Sales} = 11,032 + 821\text{Ads}$$

- Make an active prediction using these results.
 - Provide a critique as to why it may be inappropriate to rely on active predictions using these results.
10. Suppose you have regressed Y on X , and the results indicate an R^2 squared of 0.02. An analyst examining the results claims that, with such a low R^2 -squared, this regression is unsuitable for making any predictions. Is this claim correct? Why or why not? (LO7)

186

QUANTITATIVE PROBLEMS connect

11. The data in *Ch6Prob111213.xlsx* contain daily information on number of visits to your firm's website (WebVisits), number of Yahoo visitors who viewed your banner ad (YahooViews), and number of television-watchers who viewed your TV ad (TVViews). Using these data, calculate: (LO2)
- The partial correlation between:
 - WebVisits and YahooViews, controlling for TVViews

- (pCorr(WebVisits, YahooViews; TVViews))
- II. WebVisits and TVViews, controlling for YahooViews
(PCorr(WebVisits, TVViews; YahooViews))
- b. The semi-partial correlation between:
- WebVisits and YahooViews, controlling for TVViews
(pCorr(WebVisits, YahooViews; TVViews))
 - WebVisits and TVViews, controlling for YahooViews
(PCorr(WebVisits, TVViews; YahooViews))
- c. The regression equation: $\text{WebVisits} = b + m_1 \text{YahooViews} + m_2 \text{TVViews}$

Dataset available at www.mhhe.com/prince1e

12. Again using the data in *Ch6Prob111213.xlsx*, answer the following questions pertaining to the population regression equation: $\text{WebVisits} = B + M_1 \text{YahooViews} + M_2 \text{TVViews}$. (LO3)
- Test the hypothesis that M_1 is equal to zero, using a confidence level of 95%. Be sure to provide the reasoning behind your result.
 - Build a 99% confidence interval for M_2 —the population regression coefficient for TVViews. Be sure to provide the reasoning behind your result.

Dataset available at www.mhhe.com/prince1e

13. Once more, consider the data in *Ch6Prob111213.xlsx*. Assume the data-generating process can be written as: $\text{WebVisits}_i = \alpha + \beta_1 \text{YahooViews}_i + \beta_2 \text{TVViews}_i + U_i$. (LO5)
- Test the hypothesis that YahooViews has no impact on WebVisits, using a confidence level of 95%. Be sure to provide the reasoning behind your result.
 - Build a 99% confidence interval for the impact of TVViews on Webvisits. Be sure to provide the reasoning behind your result.
 - If your firm tends to offer price discounts on its website while also increasing TV advertising on Sundays, would this affect your answers

to Parts a and b?

Dataset available at www.mhhe.com/prince1e

14. The data in *Ch6Prob14.xlsx* contain information on customers' ratings of your product (CustRate), on a scale of 1 to 100, along with demographic information. The demographic information includes: income (Inc), age (Age), education (Educ), and marital status (Marr). The last variable equals one if the respondent is married and zero otherwise. Assume the data-generating process can be written as: $CustRate_i = \alpha + \beta_1 Inc_i + \beta_2 Age_i + \beta_3 Educ_i + \beta_4 Marr_i + U_i$. (LO6)
- Test the hypothesis that income has no impact on customer rating, using a confidence level of 95%. Be sure to provide the reasoning behind your result.
 - Test the hypothesis that $\beta_2 = 0.05$, using a confidence level of 90%. Be sure to provide the reasoning behind your result.
 - Build a 95% confidence interval for the impact of education on customer rating. Be sure to provide the reasoning for your result.
 - Build a 95% confidence interval for the impact of being married on customer rating. Be sure to provide the reasoning for your result.
 - Predict the change in customer rating if a customer's income increases by \$10,000, with no change in age, education, or marital status.

Dataset available at www.mhhe.com/prince1e

Basic Methods for Establishing Causal Inference

7

LEARNING OBJECTIVES

After completing this chapter, you will be able to:

- LO7.1** Explain the consequences of key assumptions failing within a causal model.
- LO7.2** Explain how control variables can improve causal inference from regression analysis.
- LO7.3** Use control variables in estimating a regression equation.
- LO7.4** Explain how proxy variables can improve causal inference from regression analysis.
- LO7.5** Use proxy variables in estimating a regression equation.
- LO7.6** Explain how functional form choice can affect causal inference from regression analysis.

dataCHALLENGE Does Working Out at Work Make for a Happy Worker?

Recently, your firm has experienced an alarmingly high rate of turnover, and management believes this is being driven by a low level of employee job satisfaction. In an attempt to improve job satisfaction, you have conducted a survey of all employees, collecting information on: job satisfaction (0–100), years of education, sex, hours per week at the company gym, and pay grade (1–5). Management is considering providing incentives for employees to exercise more during work hours, but is unsure whether doing so is likely to make a difference in employee job satisfaction.

How can you use your survey to inform this decision?

188

Introduction

In the previous chapter, we detailed the set of assumptions that will allow us to use regression analysis to assess causal relationships between variables. If these assumptions hold, the process of establishing causality and making active predictions is easy—simply estimate the regression model and apply it.

However, just because we choose to assume something doesn't make it true. There are many instances with business data, and many other data types, where the assumptions that lead to causality do not hold for the model we are estimating. Which assumptions are most susceptible to criticism? What are the consequences when they don't hold true? And perhaps most importantly, what can we do to remedy the problem when one or more of these assumptions doesn't hold true? In this chapter, we address these questions, with a focus on providing some basic, and easily implemented, solutions for the last.

Assessing Key Assumptions Within a Causal

Model

LO 7.1 Explain the consequences of key assumptions failing within a causal model.

Recall from [Reasoning Box 6.5](#) the assumptions that imply our regression estimators consistently estimate the parameters of a determining function. These assumptions are:

1. The data-generating process for an outcome, Y , can be expressed as:

$$Y_i = \alpha + \beta_1 X_{1i} + \dots + \beta_K X_{Ki} + U_i$$

2. $\{Y_i, X_{1i}, \dots, X_{Ki}\}_{i=1}^N$ is a random sample.
3. $E[U] = E[U \times X_1] = \dots = E[U \times X_K] = 0$

If these assumptions hold, then we can use our regression equation estimates as “good guesses” for the parameters of the corresponding determining function, and thus use them for causal analysis (active prediction). However, there is no guarantee that these assumptions will hold for a given dataset and population. In this section, we assess assumptions 2 and 3, describing circumstances where they are, and are not, satisfied, and explaining the consequences when they do not hold. As noted in [Chapter 6](#), these two assumptions line up closely with the two key assumptions implying causality in an experimental setting (random sample and random treatment assignment).

Why don't we assess assumption 1? This assumption may seem the most problematic, as it may appear we are simply assuming the thing we hope to estimate. However, this assumption only places mild limits on the features of the data-generating process; it leaves plenty of room for the data to tell us what it looks like. In particular, it states that the determining function is linear in the parameters, and that other factors—in the form of the error term—are additive (they simply add on at the end). In business, economic theory, and

basic accounting, well-known formulas often dictate how we express the data-generating process, and fortunately, our assumed structure is typically consistent with these formulas. As a simple example, the total costs of production that a firm faces can be written as:

$$\text{Total Costs} = \text{Fixed Costs} + f_1 \text{Factor}_1 + \dots + f_K \text{Factor}_J$$

Here, Factor_j represents a factor of production and f_j its price. If we have data on Factor₁ through Factor_K, where $K < J$, then we can write the total costs for a given firm i as:

$$\text{Total Costs}_i = \alpha + \beta_1 \text{Factor}_{1i} + \dots + \beta_K \text{Factor}_{Ki} + U_i$$

This is a way of representing the data-generating process for total costs and is consistent with assumption 1.

Of course, there are cases where assumption 1 does not hold—that is, there can be actual data-generating processes that do not have the assumed structure in assumption 1. For example, if the data-generating process is or $Y_i = \alpha + \beta_1 X_i^{\beta_2} + U_i$ or $\alpha + \beta_1 X_i \times U_i$, assumption 1 does not hold. Methods of estimating the parameters of the determining function in such cases do exist; however, these methods lie well outside the scope of this book. Although there can be cases where it may not hold, assumption 1 is reasonably defensible in a vast number of business applications. Consequently, we proceed with no further critique of assumption 1 and move to assessing the remaining two.

RANDOM SAMPLE

The second assumption we make to ensure our regression estimates are good guesses for the parameters of the determining function is that our sample is

random. Here, we consider how to draw a random sample, differentiate between random and representative samples, and then consider the consequences of using nonrandom samples.

Drawing a Random Sample There are many ways to collect a random sample, but all start with first defining the population. For example, we may define the population as all individuals in the United States, and then randomly draw Social Security numbers to build our sample. Or, we may define the population as all possible realizations for region-months over a five-year span, and treat the realizations for the region-month combinations we observed as a random sample from this population.

To elaborate further on this second example, suppose your firm operates in eight regions, and you have monthly data for all eight regions spanning five years. Hence, you have 480 observations in your sample ($12 \text{ months} \times 5 \text{ years} \times 8 \text{ regions} = 480 \text{ region-months}$). Given we are observing every region for every month during the five years, how is this a random sample, and not the entire population? The answer lies in our understanding of “possible realizations” for these region-months.

When dealing with populations that span multiple periods of time (e.g., minutes, days, months), we generally treat what was observed for a given period of time as a realization from a broader set of possibilities (what might have been). This understanding is common in sports, where we may claim that Team A would defeat Team B nine times out of ten. Here, winning and losing are possible for Team A during the period it plays Team B, but we see only one or the other actually occur. For our business example, we may be collecting information on sales for each region and month over five years. Following this same line of reasoning then, if we observed sales were 1,200 for August 2017, we would treat

190

this as a realization from a broader set of sales that were possible for August 2017 (e.g., sales could have been anywhere between 700 and 2,000 in August 2017).

Viewed in this light, assuming a random sample from data spanning

multiple periods implies that data we observe for each period are random draws from what we might have observed for those periods. This is a bit of an abstract concept, but its practical application is straightforward. For data spanning multiple periods, assuming a random sample implies the realizations we observe for each period are not consistently “special” relative to what we believe might have been observed.

To further illustrate this idea, consider U.S. airline data during the year following September 11, 2001. If we collect data at the airline-month level around this time, it is not reasonable to assume the realizations of the variables we collect (of airline revenue) are random draws from what might have occurred during those months. Rather, they likely consist of a series of extreme draws from a hypothetical population.

Random vs. Representative Sample A key merit of drawing a random sample is that, on average, it should look like a smaller version of the population from which we are drawing. We expect the information in a random sample to “represent” the information contained in its corresponding population. However, for any given sample of data, the fact that it was drawn randomly does not guarantee that it represents the population well.

Suppose your firm is interested in the relationship between an individual's age and his or her rating of the firm's newest product (on a scale of 1–100), among those visiting the firm's website. To measure this relationship, you may randomly survey 35 people on the website, asking their age and their rating of the product. In doing so, you may end up with a dataset like that in [Table 7.1](#).

Suppose from prior sampling of your customers, you were confident that approximately 30% of your customers were over the age of 40. Then, although you chose the individuals for your sample randomly (e.g., using a random number generator on your server), you may not be especially pleased with the sample that resulted. In particular, you have information on very few individuals over the age of 40 in the sample, precluding you from learning much about a significant portion of your customer base. The problem: The random sample you drew, by simple chance, is not very representative of the

population. As a result, we may worry that we can't learn very much about the preferences of our “over-40” customers.

To avoid situations like that occurring in [Table 7.1](#), it is common practice to take measures to collect a representative sample. A **representative sample** is a sample whose distribution approximately matches that of the population for a subset of observed, independent variables.

representative sample A sample whose distribution approximately matches that of the population for a subset of observed, independent variables.

The typical method for **constructing a representative sample** is as follows:

constructing a representative sample The four steps that are to be followed in building a *representative sample*.

STEP 1: Choose the independent variables whose distribution you want to be representative.

STEP 2: Use information about the population to stratify (categorize) each of the chosen variables.

STEP 3: Use information about the population to pre-set the proportion of the sample that will be selected from each stratum.

STEP 4: Collect the sample by randomly sampling from each stratum, where the number of random draws from each stratum is set according to the proportions determined in Step 3.

TABLE 7.1 Age and Rating Data for a New Product

INDIVIDUAL NUMBER	AGE	RATING
1	37	59
2	21	42
3	27	47
4	33	50

5	30	59
6	23	56
7	34	57
8	22	58
9	36	53
10	30	65
11	31	54
12	20	51
13	48	63
14	22	54
15	29	49
16	38	55
17	28	51
18	26	63
19	33	57
20	34	52
21	29	50
22	20	41
23	53	68
24	30	60
25	25	55
26	39	62
27	23	44
28	30	55
29	38	57
30	31	60
31	26	44
32	34	53
33	26	59
34	38	63
35	30	58

Let's consider this process for our age/rating example. Here, we are interested in how ratings depend on age, so we have age in the role of an independent variable. Applying the steps:

STEP 1: With just one independent variable, step 1 is trivial—we want a representative sample according to age.

STEP 2: For step 2, we need to utilize information we have about the population. Here, we know that 30% of the population is over the age of 40. Therefore, we can stratify the data into two groups: over 40, and 40 and under.

STEP 3: For step 3, we again use our knowledge of the population to determine the proportion of our sample coming from these two strata: 30% of our sample should be over 40, and 70% should be 40 and under. Thus, if our sample size is $N = 1,000$, we will have 300 who are over 40 and 700 who are 40 and under.

STEP 4: Lastly, for step 4, we may collect a random sample larger than $N = 1,000$ to ensure there are at least 300 who are over 40 and at least 700 who are 40 and under. Then, randomly select 300 from the subgroup who are over 40, and randomly select 700 from the subgroup who are 40 and under.

It is important to note that the concepts of random and representative are not mutually exclusive when it comes to data samples. For example, a sample can be both random and representative. This will happen if we have a random sample, and by the luck of the draw, it happens to be representative along one or more observed, independent variables.

However, if we construct a representative sample, then by construction, it is not truly a random sample. The question then becomes whether having a constructed representative sample, rather than a random sample, precludes us from using all of the results we've established (e.g., consistent estimators) that relied on assuming a random sample. More specifically, in constructing a representative sample, we are collecting data in a way that is not random with respect to one or more observed, independent variables. Does this type of nonrandomness confound our findings relying on the assumption of a random sample? Fortunately, as we elaborate below, the answer is "No."

Constructing a representative sample also provides valuable benefits when conducting data analysis. It ensures that we observe the pertinent range of our independent variables. As we'll discuss further in [Chapter 10](#),

observing this pertinent range makes those estimates more broadly applicable, since it broadens the values of X s that were observed (older customers in our example).

In addition, constructing a representative sample often ensures that we have substantial variation in the independent variables. While we have not explicitly detailed the calculation of the variance of our parameter estimators in regression, we note here that they are decreasing in the sample variance of the independent variables. In other words, as an independent variable, say X_1 , increases in variance in our data, the variance of our estimator for the coefficient of X_1 , $\widehat{\beta}_1$, gets smaller. This means that our inference (hypothesis tests and confidence intervals) becomes more precise.

Because constructing a representative sample has notable benefits and allows for the same basic analyses as would a random sample, it is a common and often desired practice. However, note that implementation depends on prior knowledge of the population and the ability to sample in a way that mimics the known distribution of independent variables. Hence, constructing a representative sample is not always an option in practice.

7.1 Demonstration Problem

Suppose your firm is interested in identifying determinants of its employees' job satisfaction. It wants to conduct this analysis by collecting data on a sample of its entire employee population. It has specified that it wants the sample to be representative according to employee sex and salary. The salary classification consists of just two groups: those making more than \$50,000 and those making less than \$50,000. The firm currently has 10,000 employees, of which: 3,200 are male with salary over \$50,000; 2,700 are female with salary over \$50,000; 2,300 are male with salary under \$50,000; and 1,800 are female with salary under \$50,000. The firm has a database where this information is available for each employee.

You'd like to collect a sample of 100 employees. Explain how to collect this sample so that it is representative according to sex and salary.

Answer:

Using the distribution of sex and salary for all employees, we know what percentage of our sample should come from each of the four strata: 32% should be male with salary over \$50,000; 27% should be female with salary over \$50,000; 23% should be male with salary under \$50,000; and 18% should be female with salary under \$50,000.

Consequently, after dividing the full set of employees into these four strata (using the firm's database), you should randomly sample: 32 people from the 3,200 who are male with salary over \$50,000; 27 people from the 2,700 who are female with salary over \$50,000; 23 people from the 2,300 who are male with salary under \$50,000; and 18 people from the 1,800 who are female with salary under \$50,000.

Note that, unlike the age/rating example in the text, the ability to pre-sort the population from which we are sampling into strata before sampling allows you to sample exactly N ($=100$) individuals. In contrast, for the age/rating example, we knew the distribution across strata for the population but could not pre-sort the population in advance. Therefore, we had to collect a sample that was likely larger than our desired sample size in order to make sure we had the proper number of observations in each stratum.

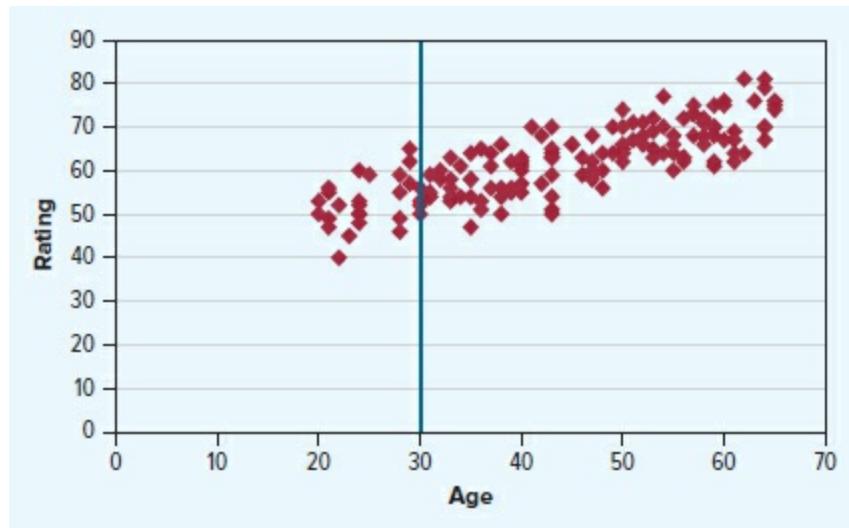
Consequences of Nonrandom Samples The construction of a representative sample generally results in a nonrandom sample. In fact, many samples utilized in business and beyond are nonrandom in one way or another. A sample that is nonrandom is also known as a **selected sample**. In this subsection, we consider two fundamental ways in which a sample can be nonrandom, or selected. It can be selected according to: (1) the independent variables (X s) or (2) the dependent variable (Y). Here, we explain the consequences of each type of nonrandomness.

selected sample A sample that is nonrandom.

Selection by Independent Variable. Consider first a data sample that is selected according to the X s. As we already noted, if we construct a representative sample as we did with our previous age/rating example, this is not random according to the X s. Or, again revisiting our age/rating example, consider a more extreme case where the program collecting the surveys retained data only on respondents under 30 years old. In both cases, we have data that are selected according to the X s—one selects 30% of the sample to have age over 40 while the other selects all of the sample to have age less than 30. Fortunately, when

194

FIGURE 7.1 Age/Rating Data Scatterplot

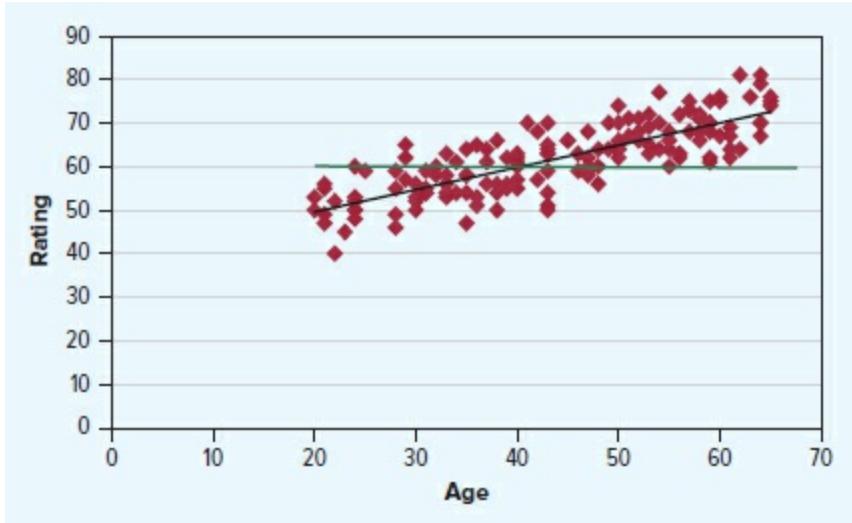


estimating a regression equation of the form $Y = \alpha + \beta_1 X_1 + \dots + \beta_K X_K$, utilizing a sample that is selected only with respect to a subset of X_1 through X_K is not consequential with regard to the consistency of the parameter estimators, hypothesis testing, or confidence intervals. In short, all reasoning boxes from [Chapter 6](#) still apply if we replace our assumption of a random sample with an assumption that the sample is selected only with respect to the independent variables (X s).

Rather than provide a formal proof of this claim, we instead provide a conceptual explanation as to why this is true. In [Figure 7.1](#), we present a scatterplot of age and rating data where $N = 150$ and we have a random sample. Now, consider the more extreme case where we instead observe only individuals who are under 30 years old. In this case, we truncate the data at Age = 30, and observe only the data points to the left of the vertical (red) line in the figure. However, using just these data points does not, per se, skew our estimate for the regression equation—in this case, a regression line. Instead, it simply limits where, along the line, we are observing data. In [Figure 7.1](#), the determining function is Rating = $40 + 0.5\text{Age}$. With a sufficient sample size, we can “see” this relationship using only data where Age is less than 30, just as we can using data with a wider Age range; if the relationship between Rating and Age is a line, we only need to see part of the line to estimate it. This same reasoning applies for the case where we force 30% of our sample to be over 40. In this case, we are altering how much data we have for different sections of the line, but this does not preclude us from properly estimating it.

We note here that, particularly in the case where we are truncating the data along X values (only observing age less than 30), this approach can have drawbacks. We may worry that the relationship between Rating and Age is different for individuals under 30 compared to those over 30. If so, by observing only ages less than 30, we could never tell whether this is the case. In short, cutoffs for the values we observe for the independent variables create what's known as an extrapolation problem, a topic we discuss in detail in [Chapter 10](#).

Selection by Dependent Variable. Now, consider a data sample that is selected according to the Y s. Unlike selection on the independent variables, selection according to the dependent variable does have consequence. Left unaddressed, it generally breaks the connection between the estimated regression equation and the population regression equation



or determining function. Considering our focus on determining functions in this book, selection on the outcome implies we generally cannot use the estimated regression equation to learn about the causal effects of the treatments.

Again, rather than provide a formal proof that selection on the dependent variable is problematic, we instead provide a conceptual explanation as to why this is true. In [Figure 7.2](#), we replicate the scatterplot of age and rating data from [Figure 7.1](#). However, we now provide a vertical cutoff. In particular, we consider the case where the sample is selected such that only observations where the rating was above 60 (above the green line) are included. In addition to the vertical cutoff at 60, we also include the actual determining function for the data-generating process that produced these data (the black line). In fact, the data-generating process that produced these data is: $\text{Rating}_i = 40 + 0.5\text{Age}_i + U_i$, where each U_i is independent (of all other U s and Age) and normal with mean zero and variance of 25. Hence this data-generating process has $E[U_i] = E[\text{Age}_i U_i] = 0$, and so with a random sample, the estimated regression line is a consistent estimator of the determining function.

We'll now show how selection on rating (the dependent variable) can cause a problem when attempting to use these data to estimate a determining function (conduct causal analysis). We'll show how selection on the

dependent variable tends to create a situation where $E[U_i] = E[X_i U_i] = 0$ may hold true for the full population, but $E[U_i] \neq 0$ and $E[X_i U_i] \neq 0$ for the selected subset of the population. That is, the errors do not have mean zero or zero correlation with the independent variables for the selected subset of the population. Consequently, when we solve the moment equations by forcing the residuals to have zero mean and zero correlation with the independent variables for the selected sample, these conditions do not match what is happening for the corresponding selected subset of the population from which the sample was drawn. Hence, our estimators no longer consistently estimate the parameters of the determining function, meaning they are no longer reliable for causal analysis.

To make this idea more concrete, consider it in the context of our rating/age example. In [Figure 7.2](#), consider the data points that lie above the cutoff of 60 for the rating but have relatively low ages (age less than 40). What do these data points all have in common? The answer is that they all have positive values for U , as can be seen from the fact that they all lie above the determining function. However, as we move to higher age levels, we can see

196

that the values for U are balanced between being positive and negative. The reason for this is straightforward—the cutoff at 60 is consequential for relatively lower age levels, since these tend to have lower ratings; however, it generally is not consequential for relatively higher age levels, since these tend to have ratings that exceed the cutoff anyway.

From these insights, we can see that the cutoff at 60 has two important consequences:

1. The mean value of the errors is positive for the selected subset.
2. The errors and age are negatively correlated.

We know consequence 1 is true since the errors have mean zero for relatively high ages but tend to be positive for relatively low ages. We know consequence 2 is true since relatively low ages tend to correspond to relatively high errors.

We conclude by noting there are corrective measures one can take to address selection issues, particularly the highly problematic issue of selection on outcomes. For example, we can make assumptions about the method of selection and adjust our estimation procedure(s) accordingly. Such estimation techniques are outside the scope of this book, but for our purposes, it is an important step to be aware of whether a sample is selected and whether the type of selection is consequential toward your findings.

NO CORRELATION BETWEEN ERRORS AND TREATMENTS

The third assumption we make to ensure our regression estimates are good guesses for the parameters of the determining function is that $E[U_i] = E[X_{1i}U_i] = \dots = E[X_{Ki}U_i] = 0$ for the data-generating process $Y_i = \alpha + \beta_1 X_{1i} + \dots + \beta_K X_{Ki} + U_i$. That is, we assume the errors have mean zero and are not correlated with the treatments in the population. As we noted when first presenting this assumption, the part that assumes the errors have zero mean is generally satisfied when there is an intercept term in the determining function. In some special cases (e.g., when there is a selected sample on the dependent variable), this part of the assumption can be violated, but we will not encounter these cases in this book.

The part of our third assumption that often comes into question is the lack of correlation between the errors and the treatments. Violation of this assumption, meaning there exists correlation between the errors and at least one treatment, is known as an **endogeneity problem**. Recall from [Chapter 6](#) that this assumption plays an analogous role in regression analysis to the assumption of random treatment assignment when conducting experiments. As is the case when there is not random treatment assignment, the existence of an endogeneity problem compromises our ability to disentangle the causal effect of a given treatment on the outcome from the effects of other, unobserved factors. We define the component(s) of the error (U_i) that are correlated with a treatment(s) (X) as **confounding factors**. This label is highly

intuitive—the existence of such factors confounds our ability to measure the causal effect(s) of one or more treatments.

endogeneity problem Correlation exists between the errors and at least one treatment.

confounding factors The component(s) of the error, U_i , that are correlated with a treatment(s), X .

How is it that confounding factors preclude us from properly measuring causal relationships between treatments and outcomes? Intuitively, the reason is that unobserved factors affecting the outcome move along with a treatment(s), and we cannot disentangle whether it is movement in the treatment(s) or these “other factors” that are causing corresponding changes to the outcome. A bit more formally, the reason links to our discussion pertaining to selection on the dependent variable. When the error and a treatment(s) are correlated in the data-generating process, then the moment equations do not mimic what is happening with the population. That is, the moment conditions force the residuals

197

to be uncorrelated with the treatment(s) in the data, but when there is an endogeneity problem, this feature of the estimated regression equation does not correspond to what is happening in the population. Consequently, we cannot expect the solution to the moment equations to produce “reliable” estimates for the actual parameters of the determining function.

To see how an endogeneity problem can impede one's ability to conduct causal analysis, consider the following example. Suppose you have been brought in as a consultant for a health and fitness company called FitU. Their primary product is the FitMaker. The company, which currently has 37 small stores throughout the United States, wants to understand how changes in price affect sales of the FitMaker. Each store manager has control over the price she charges, so there is a notable range in the price for the FitMaker across locations. FitU provides you with a cross-section of data, which includes the average price and total sales for each location in a given month.

The data are in the first three columns of [Table 7.2](#) (we'll consider the third column shortly).

Suppose we assume the following data-generating process: $\text{Quantity}_i = \alpha + \beta \text{Price}_i + U_i$. Then, we regress Quantity on Price to get our estimate for the determining function: $\text{Quantity} = 355.88 + 0.25 \times \text{Price}$. Here, we see a common problem when trying to estimate a demand equation (how quantity demanded depends on price) when the necessary assumptions do not hold. Note that our estimate implies quantity is *increasing* in price. This means, if we increase price on the FitMaker, our estimate predicts sales will go up. Of course, it is highly doubtful that this is the case; it violates one of the most fundamental laws in economics, the law of demand. This law states that quantity demanded is always decreasing in price (other things held constant); it makes good sense since we expect fewer people to buy a product as its price rises.

So, how is it that our estimates violate the law of demand, rendering them clearly inaccurate? Here, we likely have a violation of assumption 3; it is likely the case that unobserved factors affecting quantity demanded are correlated with Price (i.e., $E[\text{Price}_i U_i] \neq 0$). There are many possible confounding factors: one candidate would be local income levels. In particular, we don't expect managers to set prices for their products randomly. Instead, they observe the local demand conditions and try to set a price they believe will be most profitable. One local demand condition on which they may base their price decision is local income levels. They may choose to set a higher price when local income is high and a lower price when local income is low. If this is the case, we then have our violation, since local income likely affects quantity demanded (it is part of U), and it is correlated with the price, meaning we have $E[\text{Price}_i U_i] \neq 0$. The presence of this other factor gives the illusion that customers like higher prices. When local demand is “good” (local income levels are high), people tend to buy more of the product and the price tends to be higher, and vice versa.

Consider an alternative version of this example where we also have data on local income levels (measured as average income among households in

the region); that is, we now have the data in the fourth column of Table 7.2 as well. We can then assume a data-generating process of $\text{Quantity}_i = \alpha + \beta_1 \text{Price}_i + \beta_2 \text{Income}_i + U_i$. Then, we regress Quantity on Price and Income to get our estimate for the determining function to be: $\text{Quantity} = 13.06 - 0.40 \times \text{Price} + 0.014 \times \text{Income}$. Here, we have evidence that Income was a confounding factor. By including it in the determining function, and thus moving it out of U , it is no longer definite that our regression estimates are unreliable, as our estimates now at least satisfy the law of demand.

TABLE 7.2 Data for FitMaker

STORE NUMBER	TOTAL SALES	AVERAGE PRICE	LOCAL INCOME
1	456	588	49,499
2	583	517	56,903
3	657	764	69,066
4	480	694	53,068
5	319	364	34,390
6	382	711	45,951
7	617	576	61,721
8	642	647	63,541
9	463	516	46,257
10	414	468	41,804
11	389	498	43,052
12	331	428	35,579
13	671	661	64,894
14	445	396	40,044
15	728	554	63,723
16	559	647	58,137
17	324	278	30,724
18	356	683	43,648
19	551	481	54,456
20	424	202	35,020
21	497	536	54,428
22	433	720	52,909
23	641	503	63,337
24	409	354	41,029

25	480	215	38,342
26	340	246	32,455
27	446	271	39,936
28	553	587	58,498
29	381	493	40,216
30	498	332	42,221
31	439	491	46,356
32	502	538	52,033
33	372	615	43,983
34	438	286	39,830
35	602	419	54,892
36	515	406	46,449
37	352	566	45,642

199

We conclude by detailing the three main forms in which endogeneity problems generally materialize. These forms are not mutually exclusive—that is, an endogeneity problem can come in more than one of these forms. The three forms are:

1. Omitted variable(s)
2. Measurement error
3. Simultaneity

An endogeneity problem due to an omitted variable(s) follows the discussion of our FitU example. An **omitted variable** is any variable contained in the error term of a data-generating process, due to lack of data or simply a decision not to include it. If an omitted variable is also a confounding factor —i.e., it both affects the outcome and is correlated with a treatment(s)—then this creates an endogeneity problem. In our FitU example, we omitted local income, which was a confounding factor that affected both quantity demanded and the price. Consequently, we had an endogeneity problem due to an omitted variable.

omitted variable Any variable contained in the error term of a data-generating process, due to lack of data or simply a decision not to include it.

An endogeneity problem due to **measurement error** can arise when one or more of the variables in the determining function (typically at least one of the treatments) is measured with error. For example, suppose we want to estimate the relationship between body mass index (BMI) and daily calorie intake. We believe the data-generating process is: $BMI_i = \alpha + \beta ActCal_i + U_i$, where $ActCal$ is the actual average calorie intake for person i .

measurement error When one or more of the variables in the determining function (typically at least one of the treatments) is measured with error.

To estimate this equation, we have data on individuals' BMIs and their reported calories: $RepCal$. Note that reported calories may be inaccurate and thus may not equal actual calories. Hence, we can write reported calories as: $RepCal_i = ActCal_i + V_i$. Here, V_i represents measurement error with regard to actual calories. Using this formulation, we can rewrite our data-generating process as: $BMI_i = \alpha + \beta(ActCal_i + V_i) + (U_i - \beta V_i) = \alpha + \beta(RepCal_i) + (U_i - \beta V_i)$. Since we observe $RepCal$ and not $ActCal$, it is the determining function for this data-generating process that we must try to estimate. Note that our error term ($U_i - \beta V_i$) is generally correlated with our treatment ($RepCal_i = ActCal_i + V_i$) since both contain the measurement error (V_i). As a result, we generally have an endogeneity problem in this instance.

The third form of an endogeneity problem is simultaneity. An endogeneity problem due to **simultaneity** can arise when one or more of the treatments is determined at the same time as the outcome. Simultaneity often occurs when there exists some amount of reverse causality, where the level of the treatment depends to some extent on the realization of the outcome.

simultaneity This can arise when one or more of the treatments is determined at the same time as the outcome; often occurs when some amount of reverse causality occurs.

For example, we may be interested in whether giving an employee a raise lowers the likelihood of his leaving the following year. To assess this, we

may assume a data-generating process of $\text{Leave}_i = \alpha + \beta \text{Raise}_i + U_i$. Why might such a data-generating process suffer from an endogeneity problem? For instance, suppose the employee is planning to move out of the country the following year, and his boss was aware of this. In this case, she may refrain from giving a significant raise, knowing it will go to waste. Here, the planned move is in the error term (U) and is correlated with Raise, generating an endogeneity problem.

Endogeneity problems are pervasive. The remainder of this chapter and the entirety of the next are dedicated to dealing with endogeneity problems, particularly the omitted variable form of the endogeneity problem. However, some of the solutions we present will also be useful for endogeneity problems due to measurement error and/or simultaneity (e.g., instrumental variables).

200

Control Variables

LO 7.2 Explain how control variables can improve causal inference from regression analysis.

LO 7.3 Use control variables in estimating a regression equation.

In this section, we discuss the first of several means of dealing with an endogeneity problem. We begin by defining a control variable and demonstrating how it can alleviate an endogeneity problem. We then discuss dummy variables when used as control variables, and how to use them properly. Lastly, we provide guidelines in choosing control variables.

DEFINITION AND ILLUSTRATION

In the context of linear regression, a **control variable** is any variable included in a regression equation whose purpose is to alleviate an endogeneity

problem. As noted in the previous section, a common reason for an endogeneity problem when running a regression is the presence of an omitted variable that plays the role of a confounding factor. When you include a variable in your regression equation as a control, you eliminate that variable as a possible confounding factor, and thus eliminate it as a possible cause for an endogeneity problem. Hence, another way of characterizing a control variable is as a confounding factor that is added to a determining function.

control variable Any variable included in a regression equation whose purpose is to alleviate an endogeneity problem.

When considering if and how to use controls, it is useful to start with a very simple data-generating process that relates the outcome to the treatment in which you are interested. In general, this will look like:

$$Y_i = \alpha + \beta X_i + U_i$$

A classic concern is that there is an endogeneity problem when using regression to estimate the determining function of this process, due to variables in the error term (U) that are correlated with the treatment (X). By including control variables, we “pull” variables out of U and include them as part of the determining function. For example, we may have data on two variables, C_1 and C_2 , that we believe also affect the outcome (Y). According to the data-generating process above, these two variables are contained in U . However, we can expand the assumed data-generating process to explicitly include these variables as part of the determining function. By doing so, we have:

$$Y_i = \alpha + \beta_1 X_i + \beta_2 C_{1i} + \beta_3 C_{2i} + U'_i$$

Here, U'_i is the error term after the two controls have been removed and

included as part of the determining function. Hence, if we feared there was an endogeneity problem in the original data-generating process due to C_1 or C_2 being omitted from the determining function, our new specification that explicitly includes these two variables eliminates the problem. In this scenario, C_1 and C_2 play the role of control variables, designed to allow us to properly estimate the causal effect of the treatment (X).

We can label any variable not considered one of the treatments as a control variable. However, what makes a good control? That is, what makes a control variable particularly useful toward alleviating an endogeneity problem? To be a good control, a variable must both affect the outcome *and* be correlated with the treatment. In effect, a variable is a good control if, when not included as part of the determining function, it is a confounding factor. Therefore, by including it in the determining function, we are measuring the effect

201

REASONING BOX 7.1

CRITERIA FOR A GOOD CONTROL

Suppose we have assumed the following data-generating process:

$$Y_i = \alpha + \beta_1 X_{1i} + \dots + \beta_K X_{Ki} + U_i$$

If the variable C is a confounding factor within the data-generating process, then C is a good control—that is, including it as part of the determining function helps mitigate an endogeneity problem.

Stated more explicitly, for a variable C ,

IF:

1. C affects the outcome, Y ,
2. C is correlated with at least one treatment (X_j).

THEN:

C is a good control, and its inclusion as part of the determining function can help mitigate an endogeneity problem.

of the treatment, holding the control variable fixed. The inclusion of the control variable precludes the possibility that unaccounted-for movements in this variable are confounding our ability to properly measure the effect of the treatment.

To see a control variable in practice, consider the following example. Suppose your firm has several production facilities, and you are interested in learning the relationship between worker hours used for production and the productivity of a facility. You have weekly data on production and worker hours for eight facilities spanning 20 weeks. Here, you measure production as a proportion of maximum production (if the facility is producing at 60% relative to its maximum, Production is 0.60). You measure worker hours as a proportion of maximum worker hours (if the facility is using 70% of its maximum worker hours, Hours is 0.70).

Begin your analysis by assuming the following data-generating process:

$$\text{Production}_{it} = \alpha + \beta \text{Hours}_{it} + U_{it}$$

Here, i represents a facility and t represents a week. When you regress Production on Hours, you get the regression results in [Table 7.3](#).

According to these results, increasing worker hours by 10% of the maximum (from 50% to 60%) will increase production by 1 percentage point ($10 \times 0.10 = 1$) of the maximum. However, you are concerned that there may be an endogeneity problem using this specification for the determining function. You believe that facility managers tend to increase worker hours when experiencing malfunctions with their facility's machinery, and you believe that the proportion of machinery used affects productivity. Consequently, you believe that a facility's proportion of functional machinery

is a confounding factor in your regression equation. If you have data on this variable,

202

TABLE 7.3 Regression Output for Production Regressed on Hours

SUMMARY OUTPUT					
<i>Regression Statistics</i>					
Multiple R	0.107074189				
R Square	0.011464882				
Adjusted R Square	0.00597302				
Standard Error	0.106981733				
Observations	182				
ANOVA					
	df	SS	MS	F	Signif F
Regression	1	0.023892921	0.023892921	2.08761297	0.150
Residual	180	2.06011642	0.011445091		
Total	181	2.084009341			
	Coefficients	Standard Error	t Stat	P-value	Lowe
Intercept	0.567283312	0.041322773	13.72810351	8.24787E-30	0.485
Hours	0.103317548	0.071507089	1.444857422	0.150236507	-0.03

you can add it as a control in your regression equation. Your assumed data-generating process becomes:

$$\text{Production}_{it} = \alpha + \beta_1 \text{Hours}_{it} + \beta_2 \text{Machine}_{it} + U_{it}$$

Here, Machine is the proportion of machinery that is functional. The results of this new regression are in [Table 7.4](#).

From the results in [Table 7.4](#), it appears that leaving out the functionality

of machinery affected the findings. In particular, it caused us to underestimate the importance of worker hours toward productivity. Controlling for this variable eliminated it as a potential confounding factor. Is this the only control we should add? This is a complex question, which we will revisit shortly.

DUMMY VARIABLES

When including controls for a regression equation, it is common to make use of different types of dummy variables. A **dummy variable** is a dichotomous variable (one that takes on values 0 or 1) that is used to indicate the presence or absence of a given characteristic. We utilized dummy variables earlier in this book. For example, if a dataset contains information on whether an individual is married or not, it can do so using a variable Married, equaling 0 if not married and 1 if married.

dummy variable A dichotomous variable (one that takes on values 0 or 1) that is used to indicate the presence or absence of a given characteristic.

203

TABLE 7.4 Regression Output for Production Regressed on Hours and Machine

SUMMARY OUTPUT					
Regression Statistics					
Multiple R	0.408573898				
R Square	0.16693263				
Adjusted R Square	0.157624615				
Standard Error	0.098483478				
Observations	182				
ANOVA					

	df	SS	MS	F	Signif F
Regression	2	0.34788916	0.17394458	17.93428831	7.959
Residual	179	1.73612018	0.009698995		
Total	181	2.084009341			
	Coefficients	Standard Error	t Stat	P-value	Lowe
Intercept	0.27086966	0.063853158	4.242071484	3.54703E-05	0.144
Hours	0.248824879	0.070476775	3.530593973	0.00052723	0.109
Machine	0.337828845	0.058450754	5.779717447	3.2631E-08	0.222

COMMUNICATING DATA 7.1

IS EDUCATION GOING UP IN SMOKE?

A recent research study found that young people who regularly smoke pot are substantially less likely to graduate from high school compared to those who do not. While the study did not provide formal regression results, we can conjecture as to what those may look like. For example, we may have data for many people on the years of education they completed and the average number of days per week they smoked pot as teenagers. After running a regression, we may get the following result: $\text{Education} = 15 - 0.3 \times \text{PotUse}$. If this was the finding, we might be tempted to conclude that smoking pot three more days per week will lead to approximately one year less in completed education. However, what are some confounding factors here? Could they potentially be added as controls? If so, what might be the effect on our conclusions?

As an example, the quality of local education may be a confounding factor. Suppose you could measure local education quality and include it as a control. What do you think inclusion of this variable will do to your regression results? What other confounding factors might you be able to control for?

dichotomous characteristics as well. In particular, we typically utilize dummy variables in our regression equations in lieu of categorical, ordinal, or interval variables. We first define these types of variables, and then explain how and why dummy variables are used in their place when conducting analysis.

A **categorical variable** indicates membership to one of a set of two or more mutually exclusive categories that do not have an obvious ordering. For example, a dataset on employees may contain a variable entitled “Location,” which takes on the values “New York,” “Los Angeles,” and “Chicago.” This variable indicates where, among the firm's three offices, a given employee works.

categorical variable Indicates membership to one of a set of two or more mutually exclusive categories that do not have an obvious ordering.

An **ordinal variable** indicates membership to one of a set of two or more mutually exclusive categories that do have an obvious ordering, but the difference in values is not meaningful. For example, a dataset on individuals may contain a variable entitled “Education.” This variable may take on values 1 through 5, where: 1 means less than high school, 2 means a high school degree, 3 means some college, 4 means a college degree, and 5 means a postgraduate degree. Here, there is a clear ordering, but the fact that a college degree implies a value that is two higher than a high school degree is not meaningful per se.

ordinal variable Indicates membership to one of a set of two or more mutually exclusive categories that do have an obvious ordering, but the difference in values is not meaningful.

Lastly, an **interval variable** indicates membership to one of a set of two or more mutually exclusive categories that have an obvious ordering, and the difference in values is meaningful. For example, a dataset on individuals may contain a variable entitled “Income.” This variable may take on values 1 through 11, where: 1 means income less than \$10,000; 2 means income

between \$10,000 and \$20,000; . . . ; 9 means income between \$80,000 and \$90,000; 10 means income between \$90,000 and \$100,000; and 11 means income over \$100,000. Here, we have a clear ordering, and the difference is meaningful. For example, a value of 7 implies income about \$30,000 higher than a value of 4.

interval variable Indicates membership to one of a set of two or more mutually exclusive categories that have an obvious ordering, and the difference in values is meaningful.

Even when recorded in a numerical form, categorical, ordinal, and interval variables are seldom included in a regression equation as is. Instead, the analyst creates dummy variables for each category and uses these dummy variables in the regression. Consider again the categorical variable Location. In order to control for location in a regression, we create dummy variables, one for each category. Specifically, we create a dummy variable “New York,” which equals 1 if the employee works in New York and 0 otherwise. We also create dummy variables “Los Angeles” and “Chicago” with analogous definitions. As another example, consider again the ordinal variable “Education.” Here, we create five dummy variables: “Less than H.S.” which equals 1 if Education equals 1 and 0 otherwise, “H.S.” which equals 1 if Education equals 2 and 0 otherwise, . . . , “Postgrad” which equals 1 if Education equals 5 and 0 otherwise.

We now show how dummy variables are utilized in a regression equation and why they provide a more meaningful interpretation for a variable's effect on the outcome than the original variables they represent. Suppose we have a dataset on a firm's employees that contains information on the employees' Sales, their Commission rate, and their Location, where Location is defined as above (New York, Los Angeles, Chicago). We are interested in the effect of an employee's commission rate on that employee's sales. Hence, we assume a data-generating process of:

$$\text{Sales}_i = \alpha + \beta \text{Commission}_i + U_i$$

However, we are concerned that Location is a confounding factor, as it may affect an employee's sales and the commission rate he or she is offered. Therefore, given that we have data on location, we want to include it as a control.

At first, we may think to adjust our data-generating process as:

$$\text{Sales}_i = \alpha + \beta_1 \text{Commission}_i + \beta_2 \text{Location}_i + U_i$$

However, in this case, we cannot regress Sales on Commission and Location, since Location does not take on numerical values. Instead, we include the dummy variables we created for Location as part of the determining function, rather than the Location variable itself. Specifically, we assume the data-generating process to be:

$$\text{Sales}_i = \alpha + \beta_1 \text{Commission}_i + \beta_2 \text{LosAngeles}_i + \beta_3 \text{Chicago}_i + U_i$$

Alternatively, suppose we wished to control for the education of the employee and had data on education in the form of an Education variable, as defined previously. In this case, we may assume the data-generating process to be:

$$\begin{aligned} \text{Sales}_i = & \alpha + \beta_1 \text{Commission}_i + \beta_2 \text{HS}_i + \beta_3 \text{SomeCollege}_i + \beta_4 \text{College}_i + \\ & \beta_5 \text{PostCollege}_i + U_i \end{aligned}$$

Notice that, for both examples (controlling for location and education), we excluded one of the categories for which we created a dummy variable: New York in the first case and LessThanHS in the second. The excluded dummy variable among a set of dummy variables representing a categorical, ordinal, or interval variable is called the **base group**. We exclude one of the groups for two reasons. The first is technical, as including all of the groups

generates multicollinearity. We defer discussion of multicollinearity issues to [Chapter 10](#), where we discuss them in the context of identification problems. For now, we simply note that including all of the dummy variables will make it impossible to get regression estimates.

base group The excluded dummy variable among a set of dummy variables representing a categorical, ordinal, or interval variable.

The second reason for excluding one group is more conceptual. When we control for being in a particular group, we are interested in the effect of being in one group versus another. For example, we are interested in the effect on Sales of being in Los Angeles versus being in New York. By choosing a base group, the regression coefficients for the other groups have this “relative” interpretation. For the data-generating process $Sales_i = \alpha + \beta_1 Commission_i + \beta_2 LosAngeles_i + \beta_3 Chicago_i + U_i$, β_2 represents the difference in Sales (for a given Commission) between Los Angeles and New York. Similarly, β_3 represents the difference in Sales between Chicago and New York, and if we want to compare Los Angeles to Chicago, we simply take the difference in their coefficients: $\beta_2 - \beta_3$.

In summary, using dummy variables in place of categorical, ordinal, or interval variables accomplishes (at least one of) two important tasks. First, dummy variables allow us to control for non-numerical variables (e.g., Location). Second, they free us from imposing an often unrealistic constant effect of moving “up” groups.

To further illustrate, consider again our Sales example, and suppose we wanted to control for education. Rather than create dummies for each education group, we instead include the Education variable as is, resulting in the following data-generating process:

$$Sales_i = \alpha + \beta_1 Commission_i + \beta_2 Education_i + U_i$$

7.2

Demonstration Problem

Suppose you own a chain of bars across the United States and want to get a sense for how your bars' revenues relate to your beer prices. To do so, you collect daily data on your revenues and average beer prices. Instead of just regressing revenues on prices, you decide to control for the day of week, as this is likely correlated with demand for beer and the prices you chose to charge. Consequently, you assume the data-generating process:

$$\text{Revenue}_{it} = \alpha + \beta \text{Price}_{it} + \delta_1 \text{Monday}_{it} + \delta_2 \text{Tuesday}_{it} + \delta_3 \text{Wednesday}_{it} + \\ \delta_4 \text{Thursday}_{it} + \delta_5 \text{Friday}_{it} + \delta_6 \text{Saturday}_{it} + U_{it}$$

In estimating the corresponding regression equation, you get the following results:

$$\text{Revenue} = 3124 - 512 \times \text{Price} - 317 \times \text{Monday} - 716 \times \text{Tuesday} - 612 \times \\ \text{Wednesday} + 218 \times \text{Thursday} + 952 \times \text{Friday} + 1,116 \times \text{Saturday}$$

How would you interpret the coefficients on Monday through Saturday? What is the predicted effect on revenues from taking a given bar with a given price and changing the day from Tuesday to Thursday?

Answer:

The coefficients on Monday through Saturday represent the effect of taking a given bar with a given price and changing the day from Sunday to one of those respective days. For example, the coefficient of -716 on Tuesday implies that for a given bar with a given beer price, moving from Sunday to Tuesday lowers revenue by \$716. Here, Sunday is playing the role of our base day. If we take a given bar with a given price and change the day from Tuesday to Thursday, revenue changes by $218 - (-716) = \$934$.

Since Education is a numerical variable, we could run the corresponding regression and get an estimate for β_2 . However, consider what β_2 implies. If, say, β_2 were equal to 15, it implies that an increase in education from high school to college would increase sales by 30 ($15 \times (4 - 2)$); it also implies that an increase in education from some college to a postgraduate degree would increase sales by 30 ($15 \times (5 - 3)$). Here, we are forcing two very different changes in education to have the same effect on the outcome, which is difficult to believe in many instances, including this one. In contrast, by using dummy variables for each education category, we allow for the (likely) possibility that a change in education from a high school degree to a college degree has a different effect on sales compared to a change in education from some college to a postgraduate degree.

To conclude, we note that dummy variables, while often used as controls, need not be used exclusively in this role. In some cases, the categorical, ordinal, or interval variable they represent is the treatment whose effect we wish to measure. In such cases, the process and interpretation is still just as we described. In fact, as we elaborate further, all controls must be able to be interpreted as treatments themselves, even if they are not treatments in which we are particularly interested.

SELECTING CONTROLS

The use of controls can be crucially important when trying to eliminate endogeneity problems. However, it is common to be confronted with datasets containing many variables that might serve as controls, and be forced to decide which variables to actually utilize as controls. There is no universally accepted, algorithmic approach for selecting controls when estimating treatment effects, but there are important guidelines one should consider in doing so. We outline these guidelines and their underlying reasoning here.

The first guideline in selecting controls is theory. Any control you may consider adding to the assumed data-generating process should have a reasonable theoretical relationship with the dependent variable. If sales for

U.S. McDonald's restaurants is the dependent variable, we would not include weather in Tanzania as a control even if we had such data. This is because any explanation as to why weather in Tanzania would, in theory, affect sales at U.S. McDonald's restaurants would strain credulity. In contrast, local income is a control one may consider adding, as there are strong theoretical (economic) reasons why local income might influence the sales of a product.

Using theory as a basis for selecting controls is highly important, since it helps avoid fishing for results. If we conduct our analysis by trying unrestrained combinations of all variables at our disposal from a given dataset, it is often the case that we can attain widely varying estimates for the effect of the treatment(s). Given this, it can be tempting to simply try different combinations until a desired result is found. The requirement of a theoretical justification for adding a control can help prevent this type of fishing and thus add credibility to the findings.

In practice, when using theory as a guide, there will be two groups of variables that can be considered as possible controls: those that *should* affect the outcome, and those that *might* affect the outcome. The former group consists of variables for which theory clearly suggests a relationship between the variable and the outcome. In our McDonald's example, local income would fall in the “should” group: Individuals’ incomes should, in theory, affect their demand for a given product, either by providing more spending power to buy more of that product (for normal goods) or providing more spending power to buy substitutes (for inferior goods). In contrast, the percentage of the local population that is male may fall in the “might” group: We could make a theoretical case as to why this would affect sales, but it would not be definitive.

As a general rule, the variables that theory says should affect the outcome should all be included in the regression. The reason is twofold: First, theory strongly suggests all these variables belong as part of the data-generating process. Second, these variables also can serve as valuable data sanity checks. A **data sanity check for a regression** is a comparison between the estimated coefficient for an independent variable in a regression and the value for that coefficient as predicted by theory. If we were estimating the

effect of price on quantity demanded for gasoline, the inclusion of an income variable can serve as a simple sanity check. We know gasoline is a normal good, so as income goes up, demand for gasoline should increase. Failure to find a positive relationship between income and quantity demanded for gasoline would raise red flags about the data and/or the analysis being conducted. In contrast, finding a positive relationship between these variables gives some reassurance for the quality of the dataset and analysis.

data sanity check for a regression A comparison between the estimated coefficient for an independent variable in a regression and the value for that coefficient as predicted by theory.

208

For the variables that theory says *might* affect the outcome, how do we decide which to include and which to discard? Also, is there any reason not to include all of them? For this set of variables, we must weigh the value of their inclusion versus the value of their exclusion. For starters, we certainly want to include “good” controls (as described in [Reasoning Box 7.1](#)). That is, we want to include controls that affect the outcome and are correlated with at least one treatment. As we've discussed, these controls help alleviate endogeneity problems.

However, we should not limit our control selection to just those that can help with endogeneity. In fact, there is value in including any variable that proves to affect the outcome (any variable that has a non-zero coefficient in the determining function). Including variables that affect the outcome tends to improve the precision of our estimates for other parameters of the determining function. This can be shown using the mathematical formulas for the variance of the estimators; however, a conceptual understanding is perhaps more useful. If a variable affects the outcome, its inclusion reduces the number of unknown factors (in the error) that affect the outcome. With fewer unknown factors, we can estimate the effects of the observed factors with less uncertainty, since the data-generating process overall has less “noise.”

Checking whether a variable affects the outcome is something we've already illustrated in [Chapter 6](#). This process simply involves testing whether a variable's coefficient in the determining function is zero. If we reject the coefficient being zero, then we conclude that variable does affect the outcome and should include it as a control.

While including variables that affect the outcome as controls may alleviate endogeneity problems and likely improve precision, including variables that do not affect the outcome—called **irrelevant variables**—can be detrimental, in the form of lost precision. This again can be shown using the mathematical formulas for the variance of the estimators; however, a conceptual understanding again is perhaps more useful. Including irrelevant variables as part of the determining function is costly to precision particularly when they are correlated with a treatment(s). In this case, the co-movement between the irrelevant variables and the treatment(s) makes it more difficult to pin down the actual effect of the treatment(s) and disentangle this from the noneffect of the irrelevant variables. The estimators for the treatment will still be consistent even with irrelevant variables included, but confidence intervals for their associated parameters will be wider (i.e., less precise).

irrelevant variables Variables that do not affect the outcome.

To summarize, when selecting controls:

1. Identify variables that theoretically should or might affect the outcome.
2. Include variables that theoretically should affect the outcome.
3. For variables that theoretically might affect the outcome, include those that prove to affect the outcome empirically through a hypothesis test.
4. For variables that theoretically might affect the outcome, discard those that prove irrelevant through a hypothesis test.

One cautionary note is in order when choosing to discard seemingly irrelevant variables. Even if a variable appears not to affect the outcome after a hypothesis test (i.e., you cannot reject that its effect is zero), it is good practice to check whether the estimated effect(s) of the treatment(s) notably change when removing it from the determining function. It is possible that a

variable has little relevance toward the outcome (small enough that you fail to reject that it's zero) but is so strongly correlated with the treatment(s) that it manages to still create a notable endogeneity problem.

209

7.3 Demonstration Problem

Consider the regression results in [Table 7.5](#) for the general case where you are trying to measure the effect of a treatment on an outcome. Given these results, which controls would you include when performing a final analysis and why?

TABLE 7.5 Regression Output for Y Regressed on a Treatment, X_1 , X_2 , and X_3

SUMMARY OUTPUT					
<i>Regression Statistics</i>					
Multiple R	0.479985911				
R Square	0.230386474				
Adjusted R Square	0.222651665				
Standard Error	103.486047				
Observations	403				
ANOVA					
	df	SS	MS	F	S
Regression	4	1275942.062	318985.5156	29.7856696	1
Residual	398	4262326.042	10709.36191		
Total	402	5538268.104			
	Coefficients	Standard Error	t Stat	P-value	L
Intercept	221.3238596	26.25795869	8.428829606	6.46688E-16	1
Treatment	4.040862782	0.450382353	8.972071735	1.15913E-17	3
X1	-2.216821634	0.38749154	-5.720954921	2.08487E-08	-

X2	2.190661356	0.635192856	3.448812964	0.000623202	0
X3	-0.027136379	0.077946126	-0.348142754	0.727917042	-

Answer:

The three candidate controls are variables X_1 , X_2 , and X_3 . Presumably all three have been included because they at least may have a theoretical effect on the outcome. Of these, we should certainly retain X_1 and X_2 , as both appear to have a strong relationship with the Outcome for virtually any level of confidence (90%, 95%, 99%), as both p -values are below 0.01. We may want to exclude the variable X_3 , as its p -value is 0.728. This means we would not reject its coefficient being zero for any realistic confidence level.

It may seem obvious to exclude X_3 ; however, there are two important considerations before finalizing this decision. First, we must ask whether theory indicates X_3 should have an effect on the Outcome; if so, we should leave it in regardless of its minimal effect. Second, we should observe any changes in the coefficient for the Treatment after re-estimating the model with X_3 excluded. If the estimated effect of the Treatment is notably changed with X_3 's removal, we should keep it in.

210

Proxy Variables

LO 7.4 Explain how proxy variables can improve causal inference from regression analysis.

LO 7.5 Use proxy variables in estimating a regression equation.

Often when conducting regression analysis, we are concerned that there exists a confounding factor, but we do not have data on it. Consequently, including the confounding factor as a control to alleviate an endogeneity

problem is not an option. In such a circumstance, another remedy to consider is the use of a proxy variable. A **proxy variable** is a variable used in a regression equation in order to proxy for a confounding factor, in an attempt to alleviate the endogeneity problem caused by that confounding factor.

proxy variable A variable used in a regression equation in order to proxy for a confounding factor, in an attempt to alleviate the endogeneity problem caused by that confounding factor.

For example, suppose you own a firm that has many employees engaged in sales. You are interested in establishing whether the number of years of education an employee has affects his or her sales performance. To answer this question, you may collect data on Sales and years of Education for your employees and assume the following data-generating process:

$$\text{Sales}_i = \alpha + \beta \text{Education}_i + U_i$$

However, you are concerned that this data-generating process does not satisfy a critical assumption toward establishing causality. In particular, you worry that Education is correlated with the error term. A likely source of this correlation is through cognitive ability. That is, people with higher cognitive ability may generate more sales and also attain more years of education. In this case, cognitive ability is part of the error term, playing the role of a confounding factor.

Since cognitive ability also affects Sales, we can update our assumed data-generating process to be:

$$\text{Sales}_i = \alpha + \beta_1 \text{Education}_i + \beta_2 \text{CogAbil}_i + U_i$$

where CogAbil is employee i 's cognitive ability. Ideally, we would simply regress Sales on Education and CogAbil, using CogAbil as a control variable, thus eliminating the endogeneity problem. However, we likely do not have

data on employees' cognitive ability. Nevertheless, we may have data on another variable that could proxy for cognitive ability. In this case, we may have data on, say, employees' scores on a cognitive test the firm administers to all applicants. These scores certainly are not perfect measures of cognitive ability; however, they could proxy for it.

What makes a variable a “good” proxy variable? There are several considerations:

- First and foremost, the proxy variable must be correlated with the variable for which it is proxying.
- Second, the variables in the determining function (which includes the variable(s) being proxied) and the proxy variable are uncorrelated with the error term. This means that, if we had information on all the variables in the determining function, and thus did not need a proxy variable, then there would not be an endogeneity problem. It also means that the proxy variable provides no further information about the dependent variable beyond what is contained in the specified determining function.
- Last, the proxy variable and all other observed independent variables are uncorrelated with other factors, besides the proxy variable, that determine the *proxied* variable. This means that the proxy variable “captures” all of the correlation between the proxied variable and the treatment, and thus is able to alleviate the endogeneity problem.

211

The criteria for a good proxy variable can seem a bit technical, so it can be particularly helpful to assess these criteria, and then illustrate how to use a proxy variable, through our example. We want to use Score as a proxy variable for CogAbil—that is, we want to use employees’ test scores as a proxy for their cognitive ability. Let's look at Score as a proxy variable using the three considerations cited previously:

- For Score to be a good proxy for CogAbil, we need them first to be correlated. That is, if we estimated the parameters of the data-generating

process $\text{CogAbil}_i = \gamma + \delta \text{Score}_i + V_i$ using regression, we would get a non-zero estimate for δ .

- Second, we need Education, CogAbil, and Score all to be uncorrelated with U . That is, education, cognitive ability, and test score must all be uncorrelated with other factors that determine sales.
- Last, we need Score and Education to be uncorrelated with V ; we need each employee's test score and education to be uncorrelated with other factors (besides test score) that determine their cognitive ability.

Suppose we believe all three criteria hold. Then, we can revisit the data-generating process, plugging in for CogAbil:

$$\text{Sales}_i = \alpha + \beta_1 \text{Education}_i + \beta_2 \text{CogAbil}_i + U_i$$

And,

$$\text{CogAbil}_i = \gamma + \delta \text{Score}_i + V_i$$

So,

$$\text{Sales}_i = (\alpha + \beta_2 \gamma) + \beta_1 \text{Education}_i + \beta_2 \delta \text{Score}_i + (U_i + \beta_2 V_i)$$

Notice that, by satisfying the criteria, Score and Education are not correlated with U or V , meaning there is not an endogeneity problem when substituting Score for CogAbil. Hence, by using Score in place of CogAbil, we are able to get a consistent estimate for the effect of Education, making Score a “good” proxy variable for cognitive ability. We summarize this reasoning in [Reasoning Box 7.2](#).

CRITERIA FOR A GOOD

REASONING BOX 7.2

PROXY VARIABLE

Assume the following data-generating process:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_K X_{Ki} + U_i$$

Suppose you observe Y, X_1, X_3, \dots, X_K , but do not observe X_2 . Suppose also you observe another variable, P .

212

IF:

1. $X_{2i} = \gamma + \delta P_i + V_i$ has a non-zero value for δ (i.e., X_2 and P are correlated)
2. X_1, X_2, \dots, X_K, P are all uncorrelated with U
3. X_1, X_3, \dots, X_K, P are all uncorrelated with V

THEN:

The estimators for the coefficients for the independent variables X_1, X_3, \dots, X_K that are calculated when regressing Y on X_1, P, X_3, \dots, X_K are consistent estimators for $\beta_1, \beta_3, \dots, \beta_K$, respectively.

In words, if the three criteria for P to be a proxy variable for X_2 hold, then we get consistent estimates for the causal effects of X_1, X_3, \dots, X_K by regressing Y on X_1, P, X_3, \dots, X_K .

COMMUNICATING DATA 7.2

DOES GDP GROWTH PROXY ECONOMIC CLIMATE?

When trying to estimate the demand for a given product, a rather abstract factor that likely affects both quantity demanded and price(s) charged can be called “economic climate.” We might believe the true data-generating process

for a product looks as follows:

$$Q_{it} = \alpha + \beta_1 \text{Price}_{it} + \beta_2 \text{EconomicClimate}_{it} + U_{it}$$

Here, the unit of observation might be a state-month in the United States. Of course, there generally isn't going to be a variable we can collect that fully measures "economic climate" in a satisfying way. However, it is quite common to use proxy variables instead. A popular choice for a proxy variable for economic climate involves information about local gross domestic product (GDP). In this case, we may use information about the state's GDP in that month as a proxy for the economic climate of that state. In particular, we may include GDP growth as a proxy variable, thus regressing quantity on price and GDP growth.

When does using GDP growth as a proxy for Economic Climate allow us to get a consistent estimate for the effect of price? We need:

- GDP growth to be correlated with Economic Climate.
- Price, Economic Climate and GDP growth to be uncorrelated with "other factors" (besides Price and Economic Climate) affecting quantity demanded for the product.
- Price and GDP growth to be uncorrelated with "other factors" (besides Price) affecting Economic Climate.

We may worry that not all of these conditions hold. Price may still be correlated with other factors affecting quantity demanded such as number of local competitors. Does this mean we should not use GDP growth as a proxy? Not necessarily; rather, it suggests we may want to look for additional controls that will allow the above conditions to hold. For instance, we may try to collect data about local competition, and include a control for local competitors.

Form of the Determining Function

LO 7.6 Explain how functional form choice can affect causal inference from regression analysis.

Thus far, all of the examples we have analyzed have assumed determining functions that are not only linear in the parameters but also linear in the independent variables. By this, we mean that each independent variable included in the determining function enters as the variable itself and nothing more, implying the rate of change of the dependent variable with a change in any one independent variable (holding the others fixed) is constant.

Suppose your firm is training new salespeople and is trying to determine how much additional training translates into additional sales. You collect sales data for each employee's first year with the firm along with the amount of training they received (measured in hours). You begin by assuming the following data-generating process:

$$\text{Sales}_i = \alpha + \beta \text{Hours}_i + U_i$$

Here, the assumed form for the determining function implies that Sales change with Hours at a constant rate of β . So, if $\beta = 12$, then each increase in Hours by one causes an increase in Sales of 12. We may be concerned that this assumed structure of the relationship between Sales and Hours (it's linear) does not reflect well the true nature of how these variables relate. We may suspect that Hours affect Sales in a nonlinear way, such that they have a notably large effect for the first few Hours, but the effect diminishes as Hours becomes large. A linear determining function is unable to capture such a change in the effect of Hours on Sales, but a quadratic determining function can.

We may believe the true causal relationship between Sales and Hours looks as follows:

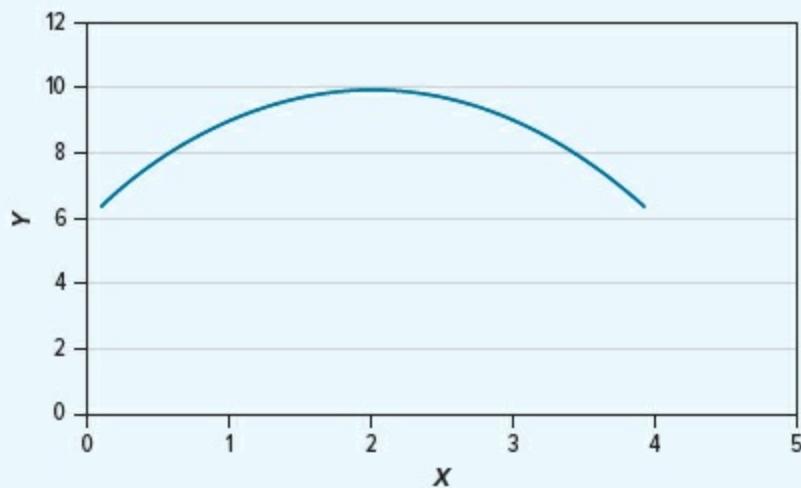
$$\text{Sales}_i = \alpha + \beta_1 \text{Hours}_i + \beta_2 \text{Hours}_i^2 + U_i$$

Assuming this relationship still allows us to use linear regression, since Sales is still linear in the parameters. We can simply set $\text{Hours} = X_1$ and $\text{Hours}^2 = X_2$, and it looks like a generic multiple regression equation.

What are the consequences of assuming a linear relationship ($\text{Sales} = \alpha + \beta \text{Hours}$) when the true causal relationship is quadratic ($\text{Sales} = \alpha + \beta_1 \text{Hours} + \beta_2 \text{Hours}^2$)? The key issue is that, by assuming a linear relationship between Sales and Hours, we constrain the “shape” of the relationship between Sales and Hours. In [Figure 7.3](#), we illustrate this point: The curve represents a quadratic relationship, and it is clear to see that this cannot be captured well by any line.

By assuming the relationship between Y and X is linear, we are assuming the coefficients on any other functions of X (e.g., X^2) are zero. In our Sales/Hours example, assuming the relationship is linear means that we are assuming the coefficient on Hours^2 is zero. This assumption can greatly affect how we characterize the relationship between Sales and Hours. If we assume it is linear, the effect is constant (β); if we assume it is quadratic, the effect is not constant—simple calculus will show it is $\beta_1 + 2\beta_2 \text{Hours}$. Thus, for a quadratic determining function, the effect of increasing Hours by one on Sales depends on the number of Hours from which you are increasing. For example, suppose we knew $\beta_1 = 10$ and $\beta_2 = -0.5$. Then, the effect of increasing Hours by one on Sales is 5 if we are starting at 5 Hours ($10 + 2(-0.5)5$) and -10 if we are starting at 20 Hours ($10 + 2(-0.5)20$). This is a big difference, and one we could not measure if we assumed a linear determining function.

FIGURE 7.3 Quadratic Relationship between Y and X



Our Sales/Hours example is a very simple version of a general challenge: choosing the form of the determining function. Even when we add a quadratic term like Hours^2 , we are still assuming the coefficients on Hours^3 , Hours^4 , etc. are all zero. So, for any given variable we'd like to include as part of the determining function, how do we decide what functions of that variable to include (X , X^2 , X^3 , etc.)? To answer this, we must consider three features of the determining function: flexibility, precision, and exposition.

Consider again our Sales/Hours example. By including Hours^2 in the determining function, we effectively increased the *flexibility* of the determining function. As we see in [Figure 7.3](#), the quadratic function of Hours allows for a curved relationship, whereas the linear function of Hours allows for only straight lines. In short, by adding more versions of Hours as part of the determining function, we are allowing for a wider range of possible shapes that the relationship between Sales and Hours can take—that is, we are making the function more flexible in Hours. In [Reasoning Box 7.3](#) we explain in greater detail how adding more versions of a variable can create unlimited flexibility.

REASONING BOX 7.3

WHY POLYNOMIALS DO THE TRICK—THE WEIERSTRASS

THEOREM

A polynomial is a function that consists of constants and variables, where the variables' exponents are whole numbers. For example, the function $f(X) = 5 + X + X^2 + 3X^4$ is a polynomial, while $f(Y) = 4/Y + Y^{3/2}$ is not.

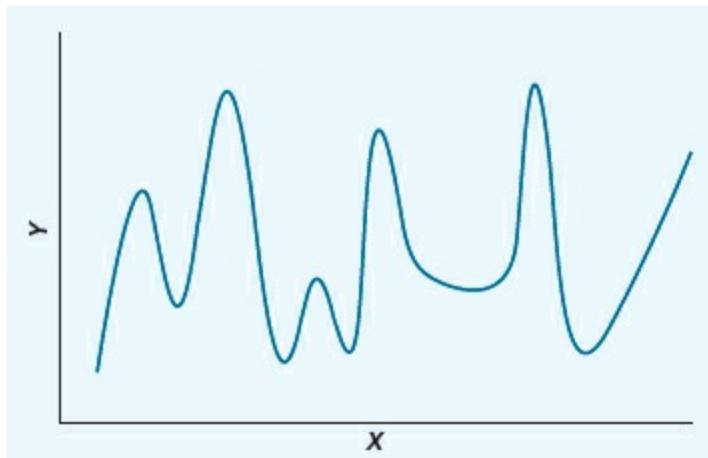
One reason polynomials are a popular choice for a determining function is their simplicity combined with their ability to generate unlimited flexibility.

What does it mean to be able to generate unlimited flexibility? To answer this, we turn to the Weierstrass approximation theorem. A simplified characterization of the theorem can be stated: If a function is continuous, it can be approximated as closely as desired with a polynomial function.

215

Consider the determining function in [Figure 7.4](#). This function clearly cannot be well approximated by a line or even a quadratic function. However, the Weierstrass theorem tells us that there is a polynomial that can get extremely close to even this highly irregular function—it just may need to contain many different powers for X (X , X^2 , X^3 , etc.). Few determining functions are as irregular as the one portrayed in [Figure 7.4](#), so large polynomials with many powers for the X s are seldom necessary to get a close approximation. This is why we seldom see polynomials with degree higher than three utilized in practice.

FIGURE 7.4 Example of a Continuous but Highly Irregular Function



COMMUNICATING DATA 7.3

TROUBLE WITH THE (LAFFER) CURVE

A famous hypothesized determining function between tax rates and tax revenues is commonly known as the Laffer curve. The *Laffer curve* is based on the idea that tax revenue will be zero both with a zero tax rate and a 100% tax rate, but is positive for tax rates in between. A simple way of capturing these features is with a quadratic function, whose maximum is somewhere between 0 and 100. While there is no widely accepted estimate for the Laffer curve, it has been used to justify claims that lowering tax rates may actually increase tax revenue (and vice versa). A possible version of the Laffer curve is in [Figure 7.5](#).

Suppose you were able to collect data containing tax rates and tax revenues across an extended period of time. If the determining function is as in [Figure 7.5](#), the data points you collect might look like those in [Figure 7.6](#).

Now, suppose you assumed a determining function of: $\text{Revenue} = \alpha + \beta\text{Rate}$. In this case, you are forcing a linear relationship when the relationship clearly is not linear—there is insufficient flexibility. Here, your estimates using a linear determining function will be especially misleading because you are likely to estimate a flat line, implying revenues do not change with tax rates. This clearly is not the case. By adding a squared version of the tax rate in the assumed determining function, you allow for the possibility of a relationship with curvature as in [Figure 7.5](#)—such flexibility is clearly crucial toward properly measuring the Laffer curve.

FIGURE 7.5 Possible Shape of the Laffer Curve

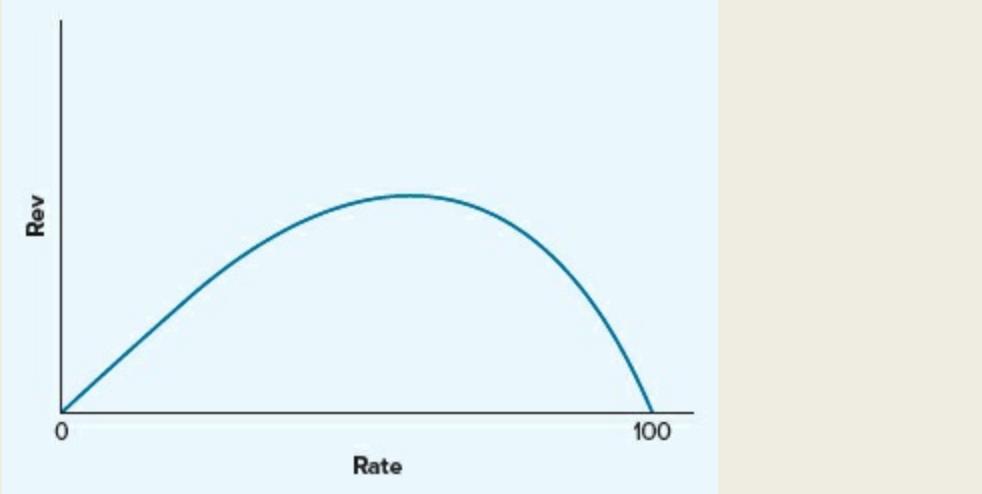
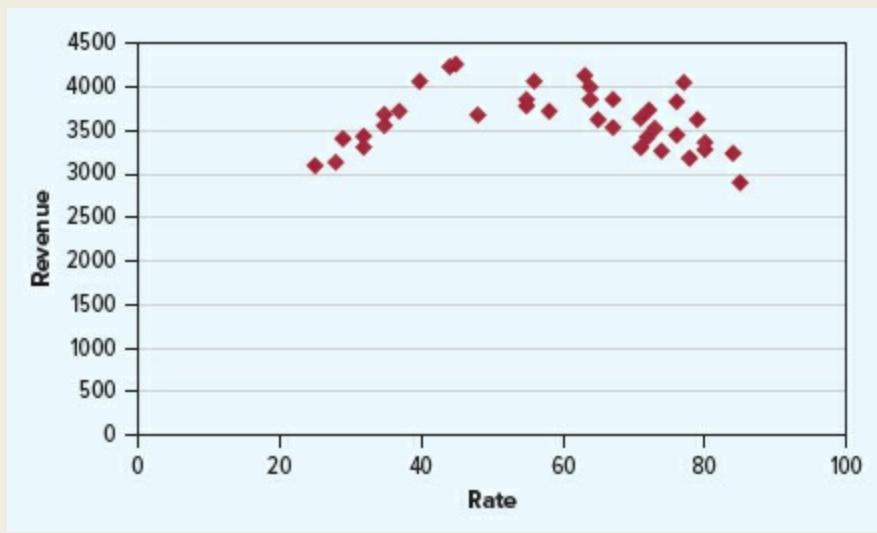


FIGURE 7.6 Possible Data on Revenue and Tax Rate



Increasing flexibility can be quite important, as we illustrated in our Sales/Hours example and in [Communicating Data 7.3](#). Consequently, it can be tempting to choose a form for the determining function that is highly flexible, including many versions of each variable (X , X^2 , X^3 , etc.). For our Sales/Hours example, we could assume, for example, the following for

$$\text{Sales} = \alpha + \beta_1 \text{Hours} + \beta_2 \text{Hours}^2 + \beta_3 \text{Hours}^3 + \beta_4 \text{Hours}^4 + \beta_5 \text{Hours}^5$$

This form for the determining function is highly flexible and allows for a wide range of shapes for the relationship between Sales and Hours. However, this high level of flexibility comes at a cost. As we add more versions of a variable to the determining function, it can reduce the *precision* of our estimates for each version. This issue is comparable to the case of including irrelevant variables. In our example, it is unlikely that Hours⁵ notably affects

217

Sales beyond the combined effects of Hours, Hours², Hours³, and Hours⁴ (meaning it is essentially irrelevant). And Hours⁵ is almost certainly correlated in some way with these other four versions. Consequently, including Hours⁵ will reduce the precision of our estimates for the effects of Hours, Hours², Hours³, and Hours⁴ in the same way the inclusion of any other irrelevant variable would.

Adding more versions of a variable not only can be costly in terms of precision, it can also adversely affect the *exposition* of the analysis. When we assume a linear relationship, explaining the relationship between the variables is straightforward—we simply report the slope of the line. To illustrate, if we assume $\text{Sales} = \alpha + \beta \text{Hours}$ and get an estimate for β of 4.6, then we would say our analysis indicates that each extra hour of training results in 4.6 more sales. In contrast, if we instead assume the determining function $\text{Sales} = \alpha + \beta_1 \text{Hours} + \beta_2 \text{Hours}^2 + \beta_3 \text{Hours}^3 + \beta_4 \text{Hours}^4 + \beta_5 \text{Hours}^5$, describing how Sales change with Hours is far more complex. In fact, the effect of an additional hour on Sales assuming this functional form is: $\beta_1 + 2\beta_2 \text{Hours} + 3\beta_3 \text{Hours}^2 + 4\beta_4 \text{Hours}^3 + 5\beta_5 \text{Hours}^4$. Even assuming the simpler quadratic relationship of $\text{Sales} = \alpha + \beta_1 \text{Hours} + \beta_2 \text{Hours}^2$ is notably more complicated to exposit, as the effect of an additional hour depends on the current number of hours (it is $\beta_1 + 2\beta_2 \text{Hours}$).

There is no universal answer to how to choose the “right” functional form. One can always choose one with many versions of the variables in it, and then test whether their coefficients are equal to zero. However, as inclusion of many versions (e.g., X , X^2 , X^3 , X^4 , X^5) typically reduces precision, it is not unusual to find that you cannot reject any of the

coefficients equaling zero, making it difficult to decide what to include and what to exclude. Common choices for the form of the determining function are linear, quadratic, and log. The first two we have already discussed; choosing between them generally comes down to whether there is ample reason to believe the effect of a variable is not constant. In what follows, we elaborate on the log functional form.

When using the *log* functional form, we mean that in the determining function, the dependent variable, an independent variable, or both are in logged form. Note that in this context, our use of log refers to the natural log, where $\log(X)$ is the exponent on the number e ($= 2.71828\ldots$) that equals X . For example, $\log(5) = 1.609$ because $e^{1.609} = 5$. The log functional form is a popular assumption for the determining function because it allows the measured effects to be interpreted as percentages. In business and elsewhere, we are often interested in the percentage change of a variable (percentage increase in sales, percentage change in profits) rather than just level changes (number of increased sales, dollar change in profits).

To see how to use log in the determining function, let's consider the three basic possibilities for using a log functional for level-log, log-level, and log-log:

- For *level-log*, the dependent variable is not in log form but an independent variable is—for example, $Y = \alpha + \beta \log(X)$.
- For *log-level*, the dependent variable is in log form but the independent variables are not—for example, $\log(Y) = \alpha + \beta X$.
- For *log-log*, both the dependent variable and an independent variable are in log form—for example, $\log(Y) = \alpha + \beta \log(X)$.

All three choices for incorporating log in the form of the determining function allow for percentage interpretations, but in different ways. We summarize how these interpretations work in [Table 7.6](#). (Note: Δ is shorthand for “change.”)

TABLE 7.6 Interpretations of ? for Different Log Functional

Forms

MODEL	DEPENDENT VARIABLE	INDEPENDENT VARIABLE	INTERPRETING β
Level-log	Y	$\log(X)$	$\Delta Y = (\beta/100)\% \Delta X$
Log-level	$\log(Y)$	X	$\% \Delta Y = (100\beta) \Delta X$
Log-log	$\log(Y)$	$\log(X)$	$\% \Delta Y = \beta \% \Delta X$

To help clarify these interpretations for log, consider again our Sales/Hours example, and the three possible ways we could incorporate log in the determining function:

- First, we may assume the determining function to be level-log, i.e.: $Sales = \alpha + \beta \log(Hours)$. For this model, if our estimate for β is, say, 400, then a 1% increase in training hours will cause an increase in sales of 4 ($= (400/100) \times 1$).
- Second, we may assume the determining function to be log-level, i.e., $\log(Sales) = \alpha + \beta Hours$. For this model, if our estimate for β is, say, 0.02, then an increase in training hours by one will cause an increase in sales by 2% ($= (0.02 \times 100) \times 1$).
- Third, we may assume the determining function to be log-log, i.e., $\log(Sales) = \alpha + \beta \log(Hours)$. For this model, if our estimate for β is, say, 0.3, then a 1% increase in training hours will cause an increase in sales by 0.3%.

This last model (log-log) is a particularly popular choice for many analyses since it measures **elasticity**, the percentage change in one variable with a percentage change in another.

elasticity The percentage change in one variable with a percentage change in another.

RISING TO THE dataCHALLENGE

Does Working Out at Work Make for a Happy Worker?

Let's return again to the Data Challenge posed at the start of the chapter: to determine if exercise hours at work affect employee satisfaction. The natural starting point is to assume the simplest data-generating process possible:

$$\text{Satisfaction}_i = \alpha + \beta \text{Hours}_i + U_i$$

However, before conducting regression analysis based on this model, we should ask whether our two key assumptions for causality are satisfied. We aren't worried about the sample, since we are surveying all employees in this case. But we should be worried about endogeneity—it is likely there are factors that affect employee satisfaction that are also correlated with the number of hours they exercise during work.

A basic step to address the endogeneity issue is to add controls. Our candidates within this dataset include: Years of Education, Sex, and Pay Grade. While it's difficult to make a conclusive argument that any of these variables should affect Satisfaction, it's easy to argue that all of them *might* affect Satisfaction. Therefore, we should consider assuming the following, more complete, data-generating process:

219

$$\text{Satisfaction}_i = \alpha + \beta_1 \text{Hours}_i + \beta_2 \text{Education}_i + \beta_3 \text{Sex}_i + \beta_4 \text{PayGrade}_i + U_i$$

Before settling on this assumed data-generating process, we should consider two more things. First, we may ask whether we believe the relationship between Hours and Satisfaction is linear—whether the effect of a change in Hours on Satisfaction is constant. For example, we may think the effect of exercise is declining, or perhaps it actually increases. Consequently, we may want to

include a quadratic term for Hours as well:

$$\text{Satisfaction}_i = \alpha + \beta_1 \text{Hours}_i + \beta_2 \text{Hours}_i^2 + \beta_3 \text{Education}_i + \beta_4 \text{Sex}_i + \beta_5 \text{PayGrade}_i + U_i$$

Second, note that PayGrade is an ordinal variable, taking on the values one through five. Therefore, we should consider using dummy variables for each pay level as part of the determining function, where:

$$\text{PayGrade1} = 1 \text{ if PayGrade} = 1, \text{ and } 0 \text{ otherwise}$$

$$\text{PayGrade2} = 1 \text{ if PayGrade} = 2, \text{ and } 0 \text{ otherwise}$$

...

$$\text{PayGrade5} = 1 \text{ if PayGrade} = 5, \text{ and } 0 \text{ otherwise}$$

We must designate one of the dummy variables as the base level (PayGrade1), leaving us with the following assumed data-generating process:

$$\text{Satisfaction}_i = \alpha + \beta_1 \text{Hours}_i + \beta_2 \text{Hours}_i^2 + \beta_3 \text{Education}_i + \beta_4 \text{Sex}_i + \beta_5 \text{PayGrade2}_i + \beta_6 \text{PayGrade3}_i + \beta_7 \text{PayGrade4}_i + \beta_8 \text{PayGrade5}_i + U_i$$

While including these added variables and added versions of variables may increase our confidence that we are able to measure a causal effect of exercise at work, we still must believe that there are not further factors influencing satisfaction that are correlated with exercise hours. If we are satisfied this is true, we are ready to generate regression estimates with causal interpretations using this model. If not, we may need to consider other methods, some of which we discuss in the next chapter.

SUMMARY

In this chapter we outlined the fundamental approaches within regression analysis that one should master in order to establish causal inference. Our

focus was on the possible existence and consequences of confounding factors, which limit or even eliminate our ability to determine the effect of a given treatment on an outcome. We explained the value of using control variables, and how to utilize dummy variables as controls. We further detailed a simple approach toward selecting the controls to add to, and remove from, the determining function. We also explained how to use proxy variables to mitigate concerns about confounding factors. We closed the chapter discussing important considerations when choosing the form of the determining function.

220

We label the topics in this chapter as “basic” not just because they are relatively easy to implement, but also because they should be understood and considered in virtually every regression analysis intended to establish causality. In the next chapter, we will discuss some relatively more advanced methods for establishing causality. These methods are a bit more complex than those discussed in this chapter, and may not be applicable in every instance. Nevertheless, they are widely applicable and can be crucial particularly when the methods discussed in this chapter are insufficient.

KEY TERMS AND CONCEPTS

base group

categorical variable

confounding factor

constructing a representative sample

control variable

data sanity check for a regression

dummy variable

elasticity

endogeneity problem

interval variable

irrelevant variable
measurement error
omitted variable
ordinal variable
proxy variable
representative sample
selected sample
simultaneity

CONCEPTUAL QUESTIONS connect

1. Suppose you have assumed the following data-generating process: $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + U_i$. You further assume that U is uncorrelated with X_1 and X_2 . For each of the four possibilities below, determine whether you can get consistent estimates of the parameters for the determining function using regression if your sample of Y , X_1 , and X_2 is selected according to: (LO1)
 - a. $X_1 > 100$
 - b. $80 < X_1 < 150$ and $40 < X_2 < 205$
 - c. $Y < 200$
 - d. $50 < X_1 < 150$ and $Y > 75$
2. Suppose you assumed the following data-generating process for Y :

$$Y_i = \alpha + \beta_1 T_i + \beta_2 X_{1i} + \beta_3 X_{2i} + \beta_4 X_{3i} + U_i$$

Suppose further that the population regression equation for T is:

$$T_i = A + B_1 X_{1i} + B_2 X_{2i} + B_3 Z_i + V_i$$

where A, B_1, B_2, B_3 are all > 0 . If you regressed Y on T only, which of the variables X_1 , X_2 , X_3 , and Z are likely to be confounding factors in your

attempt to measure the effect of T on Y ? (LO1)

3. List and define the three main forms in which endogeneity problems generally materialize. (LO1)
4. Indicate whether each of the following variables is a categorical variable, ordinal variable, or interval variable: (LO2)

221

a.

COLOR
Red
Yellow
Red
Blue

b.

PURCHASE AMOUNT
\$50–\$100
<\$50
>\$100
\$50–\$100

c.

SATISFACTION RATING (1=UNSATISFIED, 10 = HIGHLY SATISFIED)
2
4
8
5

d.

DIVISION
Electronics
Clothing
Grocery

Clothing

5. Suppose you've regressed Y on X_1, X_2, X_3 , and X_4 with the intent of measuring the causal effect of X_1 on Y . The variables X_2, X_3 , and X_4 were included as controls. After running this regression, you find that the p -value for the coefficient on X_2 is 0.23. What are reasons for keeping X_2 as a control (rather than removing it from the regression) despite this high p -value? (LO2)
6. For the following questions, suppose you've assumed the following data-generating process: $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + U_i$. Complete the following statements. (LO4)
 - a. If Z_1 is correlated with X_2 , uncorrelated with U , and uncorrelated with other factors (besides Z_1) that affect X_2 , then Z_1 may serve as a
 - b. If Z_2 is correlated with X_1 and is a component of U , then if we regress Y on X_1 and X_2 to learn the effect of X_1 on Y , Z_2 is a
 - c. If Z_3 is correlated with X_1 and is a component of U , then if we regress Y on X_1, X_2 , and Z_3 to learn the effect of X_1 on Y , Z_3 serves as a

222

-
7. You are interested in the causal effect of a variable X on an outcome Y . You regress Y on X and get the following regression result: $Y = 4.3 - 8.1X$. A colleague criticizes the lack of flexibility of your functional form choice, and insists you add X^2, X^3, X^4 , and X^5 in your regression equation. While such additions may be warranted, provide two reasons why adding these additional powers of X could be problematic. (LO6)
 8. Suppose you are an employer for a large firm that hires only candidates with undergraduate college degrees. You are interested in measuring the effect of an undergraduate degree at a top-tier university (vs. an undergraduate degree elsewhere) on first-year performance with your firm. You have data for several years on new hires' degree institutions and first-year performances. However, you believe the data-generating process for first-year performance looks like: $\text{Performance}_i = \alpha + \beta_1 \text{TopDegree}_i + \beta_2 \text{Extrovert}_i + U_i$. Here, TopDegree is a variable that equals one if the new hire got a degree from a top-tier university, and Extrovert is a

measure how extroverted the new hire is; it is believed greater extroversion leads to greater performance for this particular job. You worry that Extrovert may then be a confounding factor in a regression of Performance on TopDegree. How might you use a proxy variable to address this concern? (LO4)

9. Consider and answer the following questions. (LO5)
 - a. What is the difference between a proxy variable and a control?
 - b. Suppose you have data on Y , X_1 , and X_2 , and you are interested in the causal effect of X_1 on Y . What is the difference, if any, in the regression you would run to measure X_1 's effect on Y if you treated X_2 as a control, versus if you treated X_2 as a proxy variable?
 10. Should you add more or fewer “versions” of a variable X (e.g., X^2 , $\log(X)$, etc.) as part of the assumed functional form for a regression equation if you want improved:
 - a. Flexibility
 - b. Precision
 - c. Exposition
- Explain each answer. (LO6)

QUANTITATIVE PROBLEMS connect

Dataset available at www.mhhe.com/prince1e

11. Suppose you have collected the data in the file *Chap7 Prob11.xlsx*. These data contain information on 1,000 of your online customers pertaining to their income, household size, and the size of their purchase from your site. While the sample you collected was random, you'd like it to be representative according to income and household size. From previous surveys, you believe the distribution of the entire population of your customers to consist of: 40% with income over \$60,000 and household size less than 5; 32% with income \$60,000 or less and household size less than 5; 4% with income over \$60,000 and household size more than 4; and 24% with income \$60,000 or less and household size more than 4.

Generate a sample of 200 that is representative along these income and household size categories using this sample of 1,000. Explain how you got your new sample. (LO1)

12. The data in the file *Chap7 Prob12.xlsx* contain information on your firm's sales per capita, advertising expenditure per capita, and average local income. (LO3)

Dataset available at www.mhhe.com/prince1e

- a. Regress sales per capita on advertising expenditure per capita, controlling for local income as an interval variable, where intervals are <\$35,000, \$35,000–\$44,999, \$45,000–\$54,999, and \$55,000+, and <\$35,000 is the base group.

For the remainder of the question, assume the data-generating process is $SalesperCapita_i = \alpha + \beta_1 AdExpperCapita_i + \beta_2 Inc35-45_i + \beta_3 Inc45-55_i + \beta_4 Inc55_i + U_i$, and that all other necessary assumptions toward establishing causality and performing inference hold.

223

-
- b. Interpret the coefficients for the income intervals from your regression.
- c. According to this regression, what is the effect on sales per capita when average local income increases from \$35,000–\$44,999 to \$55,000+?
- d. Test whether there is evidence of a quadratic relationship between sales per capita and advertising expenditure per capita.
13. The data in the file *Chap7 Prob13.xlsx* contain information on an Outcome, a Treatment, and several other variables. Before analyzing the data, you note that strong theoretical arguments can be made that X_1 and X_2 affect the Outcome. Theoretical arguments also can be made that X_3 , X_4 , and X_5 affect the Outcome, but these are arguments are notably weaker. Determine the regression equation you ultimately should use to measure the effect of the Treatment on the Outcome. (LO3)

Dataset available at www.mhhe.com/prince1e

- 14.** Last year, your firm collected data on each of its 107 division managers. The data contain growth figures for each manager's division, the manager's tenure with the firm, and the manager's score on a leadership test, which was administered firmwide. These data are contained in the file *Chap7 Prob14.xlsx*. (LO5)

Dataset available at www.mhhe.com/prince1e

- a. Run a regression designed to determine the effect of manager tenure on division growth.
- b. What role, if any, can the manager's leadership test score play in the regression you ran for Part a? Explain.

Advanced Methods for Establishing Causal Inference

8

LEARNING OBJECTIVES

After completing this chapter, you will be able to:

- LO8.1** Explain how instrumental variables can improve causal inference in regression analysis.
- LO8.2** Execute two-stage least squares regression.
- LO8.3** Judge which type of variables may be used as instrumental variables.
- LO8.4** Identify a difference-in-difference regression.
- LO8.5** Execute regression incorporating fixed effects.
- LO8.6** Distinguish the dummy variable approach from a within estimator for a fixed-effects regression model.

dataCHALLENGE Do TV Ads Generate Web Traffic?

You have just signed on as an analyst for upstart comedy site [FunnyHa.com](#). In an attempt to increase traffic to the website, the advertising department of [FunnyHa.com](#) has been running ads in select counties across the United States. For each targeted county, it runs the ad 9 days per month.

FunnyHa's analytics department has purchased data on web browsing behavior for Internet users for each county across the United States. These data contain information on the county, aggregate visits for each website for each day, and average demographic information for the users in each county. The advertising department has data on when and where the ads were run, and both datasets span a 4-month period.

How can you utilize these data to determine the effect of running the ad in a county on the number of visits to the [FunnyHa.com](#) website?

225

Introduction

In the previous chapter, we outlined methods for establishing causal inference to be considered when conducting virtually any regression analysis. The methods we discussed were relatively simple and generally applicable—hence the label “basic.” In this chapter, we will introduce additional methods for establishing causal inference that we label as “advanced.” The methods we introduce involve the use of instrumental variables and methods tailored for panel datasets, focusing on fixed-effects models. These methods are generally a bit more complex both in application and explanation relative to the basic methods in the previous chapter—hence the label “advanced.” Mastering these techniques is important, because the use of basic techniques such as control variables and/or proxy variables may not be enough to establish a convincing causal link between treatment and outcome. We note that these advanced methods are not applicable in every situation. Nevertheless, they warrant

detailed discussion because they can be highly effective toward establishing causality and do apply in a wide range of settings.

Instrumental Variables

We begin this chapter with a detailed discussion of instrumental variables (defined below). It is often the case that, when attempting to establish the causal effect of one variable on another using regression, we simply do not have sufficient controls (or proxies) to accomplish the task. Rather than concede defeat, we must consider alternative ways of establishing causality using available data. Instrumental variables are almost certainly the most utilized and straightforward alternative to controls (and proxies) for establishing causality, and so they lead our discussion of advanced methods for establishing causality.

DEFINITION AND ILLUSTRATION

LO 8.1 Explain how instrumental variables can improve causal inference in regression analysis.

To better understand what instrumental variables are, it will be useful to highlight, via an example, the dilemma they are designed to solve. Consider the classic case of a firm attempting to determine how its (unit) sales depend on the price it charges for its product. Suppose this firm has many retail stores throughout the United States, and the manager of each store is given the authority to charge whatever price he or she sees fit. To estimate the effect of price on sales for this firm, you may have at your disposal a cross-sectional dataset containing information for a random sample of the firm's stores on sales and (average) price for a given quarter (e.g., July–September).

To conduct your analysis, you may begin by assuming a simple data-generating process:

$$\text{Sales}_i = \alpha + \beta \text{Price}_i + U_i$$

If we believe we have a random sample and that the unobservable factors affecting Sales are uncorrelated with Price, then we know the estimators $\hat{\alpha}$ and $\hat{\beta}$ from regressing Sales

226

on Price are consistent estimates of α and β . We have a random sample of stores, so we need be concerned only about the relationship between unobservables (U) and price. In general, as discussed in our FitMaker example in [Chapter 7](#), it is unlikely that price would be uncorrelated with other factors affecting sales, since we don't expect managers to set prices for their products randomly. Rather, they observe the local demand conditions and try to set a price they believe will be most profitable.

In our FitMaker example, we highlighted local income as a local demand condition on which managers may base their price decision. If this is the case for the firm in our current example, then assuming our simple Sales/Price data-generating process, local income is a confounding factor that would cause an endogeneity problem. We may collect data on local income and add it as a control in the determining function, resulting in an assumed data-generating process of:

$$\text{Sales}_i = \alpha + \beta_1 \text{Price}_i + \beta_2 \text{Income}_i + U_i$$

Including income in the model removes local income as a confounding factor. However, does its inclusion ensure that no other confounding factors still exist? That is, are there any other factors, other than local income, that are correlated with Price and affect Sales? Many possibilities may come to mind, including local competition, market size, and market growth rate.

This leads us to our dilemma: It may be the case that we are unable to collect data on all confounding factors or find suitable proxies. Hence, we are

unable to remove the endogeneity problem by simply including controls and/or proxy variables. Fortunately, not all is lost in such a situation, and a widely used method for measuring causality that can circumvent this problem involves instrumental variables.

In defining an instrumental variable, we begin by taking a practical but somewhat informal approach. An **instrumental variable** in the context of regression analysis is a variable that allows us to isolate the causal effect of a treatment on an outcome due to its correlation with the treatment and lack of correlation with the outcome.

instrumental variable In the context of regression analysis, a variable that allows us to isolate the causal effect of a treatment on an outcome due to its correlation with the treatment and lack of correlation with the outcome.

Before expanding this definition into something more formal, it can be helpful to discuss the intuition for why instrumental variables can be effective in establishing causality between variables. To build this intuition, consider again our Sales/Price example. As noted above, a possible confounding factor—even after controlling for local income—is local competition. The local manager may have a good sense of the amount of local competition and set price accordingly (generating a correlation between local competition and price), and the amount of local competition is almost certain to affect the firm's sales. Further, it may be difficult, as an analyst, to collect a satisfying *measure* for the amount of local competition in a given market. For example, counting the number of local competitors may not be enough, since this may not capture the intensity of competition, proximity, etc.

Our inability to control for local competition precludes us from attributing movements in Sales with Price (holding local income fixed) as the causal effect of Price on Sales. Intuitively, our problem rests in the fact that, with our current model, we are unable to observe movements in Price that do not correspond with movements in at least one other variable that also affects Sales; hence, we cannot isolate the effect of Price on Sales.

However, what if we were able separate out movements in Price in the

data that were unrelated to local competition, or other (unobserved) variables that affect Sales?

227

To illustrate, suppose we knew price differences across some of the stores were due solely to differences in fuel costs, which by themselves do not affect Sales. Then, if we analyzed how Sales moved with those particular Price differences, we might be willing to attribute this movement as the causal effect of Price; given the source of the Price variation, there is no reason to believe that other variables that affect Sales are systematically moving with these Price differences.

In our example, fuel costs can help us measure the causal effect of price because they are correlated with Price but uncorrelated with Sales. In other words, fuel costs fit our informal definition of an instrumental variable. Broadly speaking, when we have a situation where the treatment is correlated with confounding factors, finding an instrumental variable allows us to use a subset of the variation of the treatment that is not related to confounding factors to measure the causal effect of the treatment on the outcome. In our example, when two locations have different prices, we generally cannot attribute differences in Sales to Price differences (after controlling for Income), since these two locations likely differ in local competition. Consequently, rather than use all of the variation in Price across the stores to measure the effect of Price on Sales, we focus on the subset of Price movements due to variation in fuel costs. By focusing on this subset of Price movements, when two locations have different Prices only because their fuel costs differ, any difference in Sales can be attributed to Price, since fuel costs don't impact Sales per se. We illustrate this idea in [Figure 8.1](#).

Consider now a more formal characterization of instrumental variables. Suppose you've assumed the following data-generating process:

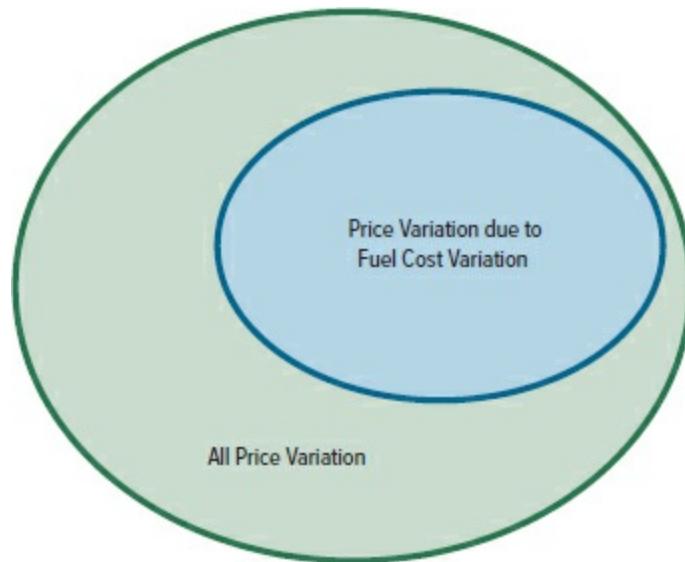
$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + U_i$$

Then, a variable, Z , is a valid instrument for X_1 if Z is both exogenous and relevant, defined as follows:

- For the assumed data-generating process, Z is **exogenous as an instrumental variable** if it has no effect on the outcome variable beyond the combined effects

exogenous as an instrumental variable A variable that has no effect on the outcome variable beyond the combined effects of all the variables in the determining function (X_1, \dots, X_K).

FIGURE 8.1 Sources of Variation in Price



of all the variables in the determining function (X_1, \dots, X_K). Put another way, Z is exogenous if the correlation between Z and the unobservables (U) is zero: $\text{Corr}(Z, U) = 0$.

- For the assumed data-generating process, Z is **relevant as an**

instrumental variable if it is correlated with X_1 after controlling for X_2, \dots, X_K . Put another way, Z is relevant if the semi-partial correlation between Z and X_1 , holding X_2, \dots, X_K constant for X_1 , is not zero: $\text{spCorr}(Z, X_1 | X_2, \dots, X_K) \neq 0$.

relevant as an instrumental variable A variable that is correlated with X_1 after controlling for X_2, \dots, X_K .

TWO-STAGE LEAST SQUARES REGRESSION

LO 8.2 Execute two-stage least squares regression.

In the previous section, we both informally and formally characterized an instrumental variable. In addition, we provided some intuition as to how an instrumental variable creates the opportunity to measure the causal effect of an endogenous variable. In this section, we detail how to perform proper analysis that allows us to utilize an instrumental variable toward measuring causal effects.

The most standard means of utilizing an instrumental variable is through **two-stage least squares regression (2SLS)**, the process of using two regressions to measure the causal effect of a variable while utilizing an instrumental variable. The conceptual underpinnings of 2SLS closely follow the intuition as to why an instrumental variable can help in measuring causal effects. Consider again our Sales/Price example, where we assumed the following data-generating process:

two-stage least squares regression (2SLS) The process of using two regressions to measure the causal effect of a variable while utilizing an instrumental variable.

$$\text{Sales}_i = \alpha + \beta_1 \text{Price}_i + \beta_2 \text{Income}_i + U_i$$

As noted before, we are concerned that Price is an endogenous variable and believe fuel costs possess the necessary characteristics (exogenous and relevant) to serve as an instrument for price. Within this framework, conceptually the first stage of 2SLS determines the subset of variation in Price that can be attributed to changes in fuel costs; we can call the variable that tracks this variation $\widehat{\text{Price}}$. Then, the second stage determines how Sales change with movements in $\widehat{\text{Price}}$. We interpret the relationship between Sales and $\widehat{\text{Price}}$ as causal since, by construction, there are no other factors influencing Sales that systematically move with $\widehat{\text{Price}}$. This means that if we see Sales correlate with $\widehat{\text{Price}}$, there is reason to interpret this co-movement as the causal effect of Price.

Now that we have laid the conceptual framework for 2SLS, we next discuss the full details of its execution. For an assumed data-generating process

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + U_i$$

suppose we believe that X_1 is endogenous and that Z is a valid instrument for X_1 . We execute 2SLS as follows. In the first stage, we assume the data-generating process:

$$X_{1i} = \gamma + \delta_1 Z_i + \delta_2 X_{2i} + \cdots + \delta_k X_{ki} + V_i$$

We then regress X_1 on Z, X_2, \dots, X_K and calculate predicted values for X_1 , defined as

$$\widehat{X}_1 = \widehat{\gamma} + \widehat{\delta}_1 Z + \widehat{\delta}_2 X_2 + \cdots + \widehat{\delta}_k X_k$$

In the second stage, we regress Y on $\widehat{X}_1, X_2, \dots, X_K$. From the second stage regression, the estimated coefficient for \widehat{X}_1 is a consistent estimate for β_1 (the causal effect of X_1 on Y).

229

The estimators for the coefficients on all other variables in the second-stage regression are also consistent estimators for their corresponding parameters in the data-generating process for Y (e.g., the estimated coefficient on X_2 is a consistent estimate for β_2).

We now apply our general discussion of executing 2SLS to our Sales/Price example. In our example, we want to instrument for Price using fuel costs assuming the data-generating process:

$$\text{Sales}_i = \alpha + \beta_1 \text{Price}_i + \beta_2 \text{Income}_i + U_i$$

To do this, we first regress Price on fuel costs and Income, assuming $\text{Price}_i = \gamma + \delta_1 \text{FuelCosts}_i + \delta_2 \text{Income}_i + V_i$. We then calculate predicted values for price from this regression, where each prediction is: $\widehat{\text{Price}}_i = \widehat{\gamma} + \widehat{\delta}_1 \text{FuelCosts}_i + \widehat{\delta}_2 \text{Income}_i$. This concludes the first stage.

In the second stage, we regress Sales on $\widehat{\text{Price}}$ and Income. The estimated coefficient on $\widehat{\text{Price}}$ is a consistent estimate for β_1 (the causal effect of Price on Sales), and the estimated coefficient on Income is a consistent estimate for β_2 (the causal effect of Income on Sales). We summarize the reasoning behind 2SLS leading to causal estimates in *Reasoning Box 8.1*.

As described above, the execution of 2SLS is quite straightforward—simply run two consecutive regressions, using the predictions from the first as an independent variable in the second. In practice, we seldom see analysts run each regression separately, for two reasons: (1) Virtually all statistical software combines this process into a single command. (2) 2SLS, as

described in [Reasoning Box 8.1](#), provides only consistent estimates; it does not ensure our ability to run hypothesis tests and build confidence intervals. It may be tempting to simply use the p -values, t -stats, etc., from the second-stage regression to perform these tasks; however, these will tend to be inaccurately measured unless corrective procedures are taken.

REASONING BOX 8.1

USING AN INSTRUMENTAL VARIABLE TO ACHIEVE CAUSAL INFERENCE VIA 2SLS

IF:

1. The data-generating process for an outcome, Y , can be expressed as:

$$Y_i = \alpha + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + U_i$$

2. $\{Y_i, X_{1i}, \dots, X_{ki}\}_{i=1}^N$ is a random sample
3. $E[U] = E[U \times X_2] = \cdots = E[U \times X_K] = 0$
4. The size of the sample is at least $30 \times (K + 1)$
5. Z is a valid instrument for X_1 (meaning it is both exogenous and relevant)

THEN:

Using Z as an instrument for X_1 , the estimated coefficients from the second stage of 2SLS are consistent estimates for $\alpha, \beta_1, \dots, \beta_K$. More specifically, let $\widehat{X}_1 = \widehat{\gamma} + \widehat{\delta}_1 Z + \widehat{\delta}_2 X_2 + \cdots + \widehat{\delta}_k X_k$ be predicted values for X_1 after regressing X_1 on Z, X_2, \dots, X_K (first stage). Then, regressing Y on $\widehat{X}_1, X_2, \dots, X_K$ generates consistent estimates for $\alpha, \beta_1, \dots, \beta_K$.

The reason that p -values and t -stats from the second stage of 2SLS are not suitable for hypothesis tests and confidence intervals is actually quite intuitive. To see this, consider again our Sales/Price example. In the second stage, we regress Sales on Price and Income, and this regression produces a p -value and t -stat associated with β_1 (the coefficient on Price). The p -value and t -stat were produced using predicted values for Price given fuel costs and Income, $\widehat{\text{Price}}$. However, they do not account for the fact that predicted values are not the same as actual values (the computer doesn't know we are using predicted values when we run the regression). Using predicted values generally comes at the expense of precision; the correct standard errors for our estimators are generally larger than the second-stage regression would report, resulting in smaller t -stats and larger p -values.

Fortunately, there is a simple solution to this problem, and as might be expected, it is to simply make a formulaic correction to the standard errors calculated in the second stage. The actual formulas are outside the scope of this book and do not provide any additional intuitive understanding of this issue. Further, they are generally embedded in the 2SLS estimation routine for virtually any statistical software. We provide an illustration of 2SLS execution, along with subsequent hypothesis testing and confidence intervals, in [Demonstration Problem 8.1](#).

8.1 Demonstration Problem

Suppose we have assumed the following data-generating process:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + U_i$$

We are concerned that X_1 is endogenous in this equation but have found an instrumental variable Z . We execute 2SLS to get estimates for the parameters of the determining function. Note that we can do this using the "Two-stage least

squares” option in the “Modeling data” toolbar in XLSTAT (an add-on for Excel), or by running the command “ivreg $Y X_2 X_3 (X_1 = Z)$ ” in STATA. (Comparable commands can be used in SAS, R, etc.)

The results of 2SLS are presented in [Table 8.1](#) below. Test whether these parameters differ from zero using 95% confidence.

TABLE 8.1 2SLS Estimates for Y Regressed on X_1 , X_2 , and X_3

DEP VAR: Y	COEF.	STD. ERR.	t - STAT	$P > t $	95% CONF. INTERVAL
X_1	-0.452253	0.2177574	-2.08	0.038	-0.8803071, -0.0241989
X_2	0.9424757	0.3443027	2.74	0.006	0.2656666, 1.619285
X_3	-1.173935	0.1723476	-6.81	0.000	-1.512725, -0.8351444
Constant	117.6513	19.10515	6.16	0.000	80.09556, 155.207

Answer:

The p -values are all less than 0.05 ($= 1 - 0.95$), indicating each of the parameters is significantly different from zero with 95% confidence. Consistent with this conclusion, none of the confidence intervals contains the number zero.

The method of 2SLS has intuitive appeal and is almost certainly the most widely understood and implemented method for utilizing instrumental variables. However, it is worth noting both as a unifying theme with our discussion of OLS and as the foundation for more advanced analysis (beyond the scope of this book but briefly introduced below) that 2SLS is just the solution to a series of sample moment equations. To illustrate, recall that, for a data-generating process $Y_i = \alpha + \beta_1 X_{1i} + \cdots + \beta_K X_{Ki} + U_i$, the associated sample moment equations for multiple regression of Y on X_1, \dots, X_K are:

$$\begin{aligned}\frac{\sum_{i=1}^N e_i}{N} &= \frac{\sum_{i=1}^N (Y_i - \hat{\alpha} - \hat{\beta}_1 X_{1i} - \dots - \hat{\beta}_K X_{Ki})}{N} = 0 \\ \frac{\sum_{i=1}^N e_i \times X_{1i}}{N} &= \frac{\sum_{i=1}^N (Y_i - \hat{\alpha} - \hat{\beta}_1 X_{1i} - \dots - \hat{\beta}_K X_{Ki}) \times X_{1i}}{N} = 0 \\ \frac{\sum_{i=1}^N e_i \times X_{2i}}{N} &= \frac{\sum_{i=1}^N (Y_i - \hat{\alpha} - \hat{\beta}_1 X_{1i} - \dots - \hat{\beta}_K X_{Ki}) \times X_{2i}}{N} = 0 \\ &\quad \dots \\ \frac{\sum_{i=1}^N e_i \times X_{Ki}}{N} &= \frac{\sum_{i=1}^N (Y_i - \hat{\alpha} - \hat{\beta}_1 X_{1i} - \dots - \hat{\beta}_K X_{Ki}) \times X_{Ki}}{N} = 0\end{aligned}$$

If we instead want to execute 2SLS where we instrument for X_1 using instrument Z , we solve the following sample moment equations:

$$\begin{aligned}\frac{\sum_{i=1}^N e_i}{N} &= \frac{\sum_{i=1}^N (Y_i - \hat{\alpha} - \hat{\beta}_1 X_{1i} - \dots - \hat{\beta}_K X_{Ki})}{N} = 0 \\ \frac{\sum_{i=1}^N e_i \times Z_i}{N} &= \frac{\sum_{i=1}^N (Y_i - \hat{\alpha} - \hat{\beta}_1 X_{1i} - \dots - \hat{\beta}_K X_{Ki}) \times Z_i}{N} = 0 \\ \frac{\sum_{i=1}^N e_i \times X_{2i}}{N} &= \frac{\sum_{i=1}^N (Y_i - \hat{\alpha} - \hat{\beta}_1 X_{1i} - \dots - \hat{\beta}_K X_{Ki}) \times X_{2i}}{N} = 0 \\ &\quad \dots \\ \frac{\sum_{i=1}^N e_i \times X_{Ki}}{N} &= \frac{\sum_{i=1}^N (Y_i - \hat{\alpha} - \hat{\beta}_1 X_{1i} - \dots - \hat{\beta}_K X_{Ki}) \times X_{Ki}}{N} = 0\end{aligned}$$

Notice that we have simply replaced the sample moment equation that generates zero correlation in the sample between the residuals and X_1 with a sample moment equation that generates zero correlation between the residuals and Z . By doing this switch, our sample moment equations now mimic what we assume is happening in the data-generating process, allowing us to infer causality (i.e., X_2, \dots, X_K , and Z are all uncorrelated with U).

These new sample moment equations yield the same solution that we get when executing 2SLS as described previously.

Thus far, our discussion of instrumental variables and 2SLS has focused on just one endogenous variable and one instrumental variable. However, this discussion easily extends into situations involving more than one of either, or both. First, consider the case where there is more than one endogenous variable in the assumed data-generating process. To be concrete, suppose that, for the data-generating process

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_2 + \cdots + \beta_K X_{Ki} + U_i$$

232

we believe X_1 and X_2 are endogenous. Then, finding a single instrument—say, Z —is not enough to solve the problem. This is easy to see if we think in terms of the sample moment equations. Specifically, we can replace one of the sample moment equations, e.g., $\frac{\sum_{i=1}^N e_i \times X_{1i}}{N} = 0$, with the sample moment equation involving Z ($\frac{\sum_{i=1}^N e_i \times Z_i}{N} = 0$). However, this requires us to retain an equation involving an endogenous variable, or else remove it and have fewer equations than estimators. Either choice is unacceptable: The first generally means our estimators are not consistent, and the second generally means we cannot find a solution.

Broadly speaking, when we have multiple endogenous variables, we need multiple instruments. If there are J endogenous variables in the determining function, we need at least J instruments. This allows us to replace every sample moment condition forcing the sample correlation between an endogenous variable and the residuals to be zero (e.g., $\frac{\sum_{i=1}^N e_i \times X_{1i}}{N} = 0$) with one that forces the sample correlation between an instrumental variable and the residuals to be zero ($\frac{\sum_{i=1}^N e_i \times X_{1i}}{N} = 0$). Further, it must be the case that the instruments are “fully correlated” with the endogenous variables.

Rather than detail what this means technically, we simply note that, roughly speaking, this means it cannot be the case that an endogenous variable is uncorrelated with any of the instrumental variables.

To see how we deal with multiple endogenous variables using 2SLS, consider again the data-generating process:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_K X_{Ki} + U_i$$

where we believe X_1 and X_2 are endogenous. In this case, we need at least two instrumental variables in order to get consistent estimates for β_1 and β_2 . Suppose we have identified two instruments, Z_1 and Z_2 , that we believe are uncorrelated with U , and are “fully correlated” with X_1 and X_2 . Then, in the first stage, we regress X_1 on $Z_1, Z_2, X_3, \dots, X_K$ and use this regression to get predicted values for X_1, \widehat{X}_1 . We do the same for X_2 by regressing X_2 on $Z_1, Z_2, X_3, \dots, X_K$ and use this regression to get predicted values for X_2, \widehat{X}_2 . In the second stage, we regress Y on $\widehat{X}_1, \widehat{X}_2, X_3, \dots, X_K$, which will yield consistent estimates for $\alpha, \beta_1, \dots, \beta_K$. As before, the solution for $\widehat{\alpha}, \widehat{\beta}_1, \dots, \widehat{\beta}_K$ when executing 2SLS is the same one we get when solving the sample moment conditions that incorporate the instrumental variables:

$$\begin{aligned}
\frac{\sum_{i=1}^N e_i}{N} &= \frac{\sum_{i=1}^N (Y_i - \hat{\alpha} - \hat{\beta}_1 X_{1i} - \dots - \hat{\beta}_K X_{Ki})}{N} = 0 \\
\frac{\sum_{i=1}^N e_i \times Z_{1i}}{N} &= \frac{\sum_{i=1}^N (Y_i - \hat{\alpha} - \hat{\beta}_1 X_{1i} - \dots - \hat{\beta}_K X_{Ki}) \times Z_{1i}}{N} = 0 \\
\frac{\sum_{i=1}^N e_i \times Z_{2i}}{N} &= \frac{\sum_{i=1}^N (Y_i - \hat{\alpha} - \hat{\beta}_1 X_{1i} - \dots - \hat{\beta}_K X_{Ki}) \times Z_{2i}}{N} = 0 \\
\frac{\sum_{i=1}^N e_i \times X_{3i}}{N} &= \frac{\sum_{i=1}^N (Y_i - \hat{\alpha} - \hat{\beta}_1 X_{1i} - \dots - \hat{\beta}_K X_{Ki}) \times X_{3i}}{N} = 0 \\
&\dots \\
\frac{\sum_{i=1}^N e_i \times X_{Ki}}{N} &= \frac{\sum_{i=1}^N (Y_i - \hat{\alpha} - \hat{\beta}_1 X_{1i} - \dots - \hat{\beta}_K X_{Ki}) \times X_{Ki}}{N} = 0
\end{aligned}$$

We now consider cases where we have more instrumental variables than endogenous variables, first focusing on the case where we have one endogenous variable and multiple instrumental variables. In our Sales/Price example, suppose in addition to fuel costs, we believed local minimum wage was related to the Price (higher wage costs lead to higher prices charged) but was unrelated to Sales. Then, we have two instrumental variables for our single endogenous variable, Price. More broadly, for an assumed data-generating process

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_K X_{Ki} + U_i$$

suppose we believe that X_1 is endogenous and that Z_1 and Z_2 are valid instruments for X_1 . In executing 2SLS, in the first stage, we regress X_1 on $Z_1, Z_2, X_2, \dots, X_K$ and calculate predicted values for X_1, \widehat{X}_1 . In the second stage, we regress Y on $\widehat{X}_1, X_2, \dots, X_K$.

We note that 2SLS is not the most efficient means of estimation when

there are more instrumental variables than endogenous variables; that is, we could get consistent estimates that are more precise (i.e., have smaller standard errors) using alternative methods. When we have more instruments than endogenous variables, we have more moment conditions than parameters for which we must solve (because we replace the moment conditions involving correlation with endogenous variables with moment conditions involving correlation with the instruments). When implementing 2SLS in this case, rather than solve all of the moment conditions (which we generally cannot do), it solves a simple minimization problem that combines all of the sample moment conditions. In particular, the solution to 2SLS is the same we get if we square each sample moment, add them up, and find the estimates that minimize this sum.

For this approach, each sample moment is treated equally, but this needn't be the case. A more advanced and efficient estimation method, called the *generalized method of moments* (often shortened to GMM) recognizes this and efficiently combines these equations (typically weighing some sample moments more than others) to reach a solution. In short, when instruments outnumber endogenous variables, 2SLS is effective—it yields consistent estimates—but alternative methods (e.g., GMM) can improve precision.

We conclude this section with a simple summary of 2SLS for the general case, where we have J endogenous variables and $L \geq J$ instrumental variables. For an assumed data-generating process

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_K X_{Ki} + U_i$$

suppose we believe that X_1, \dots, X_J are endogenous and that Z_1, \dots, Z_L are valid instruments for X_1, \dots, X_J . Execution of 2SLS proceeds as follows:

1. Regress X_1, \dots, X_J on $Z_1, \dots, Z_K, X_{J+1}, \dots, X_K$ in J separate regressions.

2. Obtain predicted values $\widehat{X}_1, \dots, \widehat{X}_J$ using the corresponding estimated regression equations in Step 1. This concludes “Stage 1.”
3. Regress Y on $\widehat{X}_1, \dots, \widehat{X}_J, X_{J+1}, \dots, X_K$, which yields consistent estimates for $\alpha, \beta_1, \dots, \beta_K$. This is “Stage 2.”

EVALUATING INSTRUMENTS

LO 8.3 Judge which type of variables may be used as instrumental variables.

Thus far, we have detailed the key characteristics of instrumental variables and how to use them in practice. We know an instrumental variable must be exogenous and relevant, and if so, we can use 2SLS to get consistent estimates for the parameters of the determining

234

function. But for any given candidate instrumental variable, can we assess whether it actually possesses these two characteristics? The remainder of this section aims to answer this question.

EXOGENEITY

Recall that an instrumental variable is exogenous if it is uncorrelated with unobservables affecting the dependent variable. This means, for a data-generating process $Y_i = \alpha + \beta_1 X_{1i} + \dots + \beta_K X_{Ki} + U_i$, an instrumental variable Z must have $\text{Corr}(Z, U) = 0$. We'd like to be able to verify that there is no correlation between Z and U , but is there a way to test this?

Before answering this question, we start by highlighting an approach that may seem capable of answering the question, but in fact is completely ineffective. In particular, we could regress Y on X_1, \dots, X_K , and calculate the residuals as:

$$e_i = Y_i - \hat{\alpha} - \hat{\beta}_1 X_{1i} - \cdots - \hat{\beta}_K X_{Ki}$$

We could then calculate the sample correlation between Z and the residuals, believing this to be an estimator for the correlation between Z and U . The problem with this approach is that the residuals were calculated using a regression with an endogenous variable (if not, then there is no need for an instrumental variable). Consequently, our parameter estimates are not consistent (we cannot trust them to be close to the actual parameters), meaning the sample correlation between Z and the residuals generally is not an estimator for the correlation between Z and U .

Unfortunately, when it comes to testing for exogeneity of instrumental variables, there are significant limitations. If the number of instrumental variables is equal to the number of endogenous variables, there is no way to test for exogeneity. For example, in our Sales/Price example where fuel costs are an instrumental variable for Price, there is no way to test whether fuel costs are exogenous. If the number of instrumental variables is greater than the number of endogenous variables, there are tests one can perform to find evidence that at least some instrumental variables are *not* exogenous, but there is still no way to test that *all* the instruments *are* exogenous.

With no tests to definitively establish exogeneity for our instrumental variables, we are left to make theoretical arguments. In our Sales/Price example, exogeneity of fuel costs implies that, after controlling for price and local income, fuel costs do not affect Sales. We cannot prove this to be true using the data. However, we can utilize knowledge of the industry, economics, and so on to make the case that this assumption likely holds. For example, suppose the product was breakfast cereal. We could argue, rather convincingly, that consumers will not take into account, in any significant way, the price they pay to fill up their cars when making cereal purchases. Higher fuel costs may affect a family's budget, but cereal is not likely to be strongly affected as compared to other more discretionary purchases, such as dining out. Fuel costs may affect the price consumers pay for cereal, but

controlling for this effect (by controlling for Price), fuel costs likely have no other discernable effect on cereal-purchasing behavior.

Of course, there can always be counterarguments to any theoretical claim of exogeneity for an instrumental variable(s). However, this situation maps well into our discussion of deductive reasoning in [Chapter 2](#). In particular, exogeneity of the instrumental variable(s) is an untested assumption that helps lead to consistent estimators. If someone does not believe the results of a 2SLS regression, then that person must disagree with at least one

235

of the assumptions leading to consistency of the estimates—and exogeneity of the instruments is often a disputed assumption. Unless there are alternative instrumental variables available on whose validity all can agree, it is left to theoretical debate as to whether the instrumental variables used are, in fact, exogenous.

Relevance Unlike exogeneity, testing for the relevance of an instrumental variable is not only possible but quite simple. In fact, we can seamlessly add a test for relevance when conducting 2SLS. Recall that, for the data-generating process

$$Y_i = \alpha + \beta_1 X_{1i} + \cdots + \beta_K X_{Ki} + U_i$$

where X_1 is endogenous, Z is relevant if it is correlated with X_1 after controlling for X_2, \dots, X_K . We can assess whether this is true by regressing X_1 on Z, X_2, \dots, X_K —exactly what we do in the first stage of 2SLS.

To illustrate, let's revisit our Sales/Price example, where we've assumed $\text{Sales}_i = \alpha + \beta_1 \text{Price}_i + \beta_2 \text{Income}_i + U_i$, and we want to use fuel costs as an instrumental variable for Price. When executing 2SLS, in the first stage we regress Price on Income and fuel costs. Suppose the regression results are as shown in [Table 8.2](#). To test for relevance of our instrumental variable (fuel costs), we want to test whether the coefficient on fuel costs differs from zero.

Using a significance level of 5%, we see (e.g., using the p -value) that we can reject the coefficient on fuel costs being zero in the population, and would thus conclude it is relevant.

It is important to note that, when testing for relevance of an instrument in the first stage of 2SLS, we are *not* testing for a causal effect of the instrumental variable (Z) on the outcome (Y). An instrument, Z , is still relevant even if the relationship with the endogenous variable, say X_1 , is purely correlational and not causal. Consequently, as was noted in [Reasoning Boxes 6.2 and 6.3](#), we can build confidence intervals and run hypothesis tests concerning partial correlations by just assuming: (1) a random sample, (2) a large sample [$30 \times (K + 1)$], and (3) homoscedasticity ($\text{Var}(Y|X) = \sigma^2$). We then test for relevance just as we test for population correlations with regression.

Testing for relevance when there are multiple instrumental variables is similar to the case with just one. For simplicity, consider a data-generating process

$$Y_i = \alpha + \beta_1 X_{1i} + \cdots + \beta_K X_{Ki} + U_i$$

where X_1 is endogenous and we are considering Z_1 and Z_2 as instrumental variables for X_1 . Again, we test for relevance using the first-stage regression, where we've regressed X_1 on Z_1 , Z_2 , X_2 , ..., X_K . Just as with a single instrumental variable, we can test whether either of the coefficients on Z_1 and Z_2 differs from zero by, for instance, looking at their p -values. In

TABLE 8.2 Regression Output for Price Regressed on Income and Fuel Costs

	COEFFICIENTS	STANDARD ERROR	t STAT	P-VALUE	LOW
Intercept	47.4377936	17.77488713	2.668809835	0.007853104	12.51
Income	-0.239175653	0.274559067	-0.871126403	0.384092613	-0.77

Fuel Costs	0.414024394	0.109815985	3.77016508	0.000182116	0.198
------------	-------------	-------------	------------	-------------	-------

236

principle, we could establish that at least one of Z_1 and Z_2 has a non-zero partial correlation with X_1 by using what's known as a joint test of significance. However, using a joint test allows for the possibility that we conclude at least one of Z_1 and Z_2 has a non-zero partial correlation with X_1 despite neither indicating a non-zero correlation when tested individually. Such an approach is valid but less convincing.

In general, it is important to establish convincing evidence that an instrumental variable(s) is relevant, as doing so avoids common criticisms of instrumental variables centered on the usage of weak instruments. A **weak instrument** is an instrumental variable that has little partial correlation with the endogenous variable whose causal effect on an outcome it is meant to help measure. For an instrument to be deemed weak, it need not have zero partial correlation with its associated endogenous variable. But it typically has a small enough partial correlation so as to fail to reject a zero partial correlation when testing for relevance. Using a weak instrumental variable(s) generally results in very poor precision in the second-stage estimates of 2SLS. This is to be expected, since it is movement of the endogenous variable (X_1) with the instrumental variable (Z) that we are using to measure the effect of the endogenous variable on the outcome (Y). If X_1 and Z have very little co-movement, we are left with little variation in X_1 to learn its effect on Y .

weak instrument An instrumental variable that has little partial correlation with the endogenous variable whose causal effect on an outcome it is meant to help measure.

8.2

Demonstration Problem

Suppose we have assumed the following data-generating process:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + U_i$$

We are concerned that X_1 is endogenous in this equation, but have found two instrumental variables Z_1 and Z_2 . We execute 2SLS to get estimates for the parameters of the determining function. Note again that we can do this using the “Two-stage least squares” option in the “Modeling data” toolbar in XLSTAT (an add-on for Excel), or by running the command “ivreg Y X₂ X₃ (X₁ = Z), first” in STATA. The addition of “first” in STATA will present the results of the first and second stage regressions.

The results of the first stage of 2SLS are presented in [Table 8.3](#), and the results of the second stage of 2SLS are presented in [Table 8.4](#). Test whether either of the instruments is individually relevant using a confidence level of 95%. Then, comment on how your findings for relevance relate to your estimate for the effect of X_1 in the second stage.

237

TABLE 8.3 Regression Results for X_1 Regressed on X_2 , X_3 , Z_1 , and Z_2

DEP VAR: X_1	COEF.	STD. ERR.	t-STAT	P > t	95% CONF. INTERVAL
X2	0.006838	0.1185558	0.06	0.954	-0.2260519, 0.239728
X3	0.0507664	0.1008174	0.50	0.615	-0.1472785, 0.2488113
Z1	-0.106104	0.1604313	-0.66	0.509	-0.421254, 0.2090459
Z2	-0.1465518	0.1128026	-1.30	0.194	-0.3681401, 0.0750366
Constant	19.22488	6.617677	2.91	0.004	6.225176, 32.22459

TABLE 8.4 Regression Results for Y Regressed on \widehat{X}_1 , X_2 , and X_3

DEP VAR:	COEF.	STD. ERR.	t-	P > t	95% CONF.
----------	-------	-----------	----	--------	-----------

Y			STAT		INTERVAL
X1	-0.9102377	0.4883163	-1.86	0.063	-1.869478, 0.0490026
X2	1.224836	0.0846242	14.47	0.000	1.058602, 1.39107
X3	1.777996	0.0761058	23.36	0.000	1.628495, 1.927497
Constant	38.61527	5.744695	6.72	0.000	27.33049, 49.90006

Answer:

Using the p -values in the first stage (Table 8.3), we see that both are well above 0.05 (0.509 and 0.194, respectively). Therefore, we fail to reject that the partial correlation between X_1 and Z_1 , and between X_1 and Z_2 , is zero.

The weak correlation in the first stage manifests in the second. Here, we see that, while we are able to measure the effects of X_2 and X_3 on Y very precisely, we have a rather wide confidence interval for the effect of X_1 by comparison.

The effect of X_1 appears to be similar in magnitude (in absolute value) compared to the effects of X_2 and X_3 , but because it is measured noisily, we cannot reject (with 95% confidence) the possibility that its effect in the population is actually zero (p -value of $0.063 > 0.05$).

CLASSIC APPLICATIONS OF INSTRUMENTAL VARIABLES FOR BUSINESS

When facing the prospect of an endogenous variable in regression analysis, the implementation of instrumental variables techniques (with the exception of establishing exogeneity) is a matter of following procedure. However, identifying variables that may be effective instrumental variables often requires at least a touch of creativity. While it is not plausible for us to provide a general, step-by-step process for discovering valid instruments, we can highlight some good types of variables to consider for common business applications. To this end, we highlight two variable types: cost variables and policy changes.

Cost variables are popular choices as instrumental variables, particularly in demand estimations. Firms often want to learn the determinants of the

number of units sold (or services demanded), including price and other factors. They will assume a data-generating process that looks like:

$$\text{Quantity}_i = \alpha + \beta_1 \text{Price}_i + \beta_2 X_{1i} + \cdots + \beta_{K+1} X_{Ki} + U_i$$

In general, this formulation raises the concern that Price is correlated with unobservables affecting Quantity and so is endogenous.

If we choose to address this issue using instrumental variables, cost variables often prove valid. In particular, any variable that affects the costs of producing the good or service (input prices, cost per unit, etc.) can prove to be a valid instrument for Price. This is because prices charged typically depend on costs—almost always for per-unit costs, and even sometimes for fixed costs—and costs typically do not affect the demand for a product or service outside their impact on price. Hence, cost variables are often both relevant and exogenous when used to instrument for price in a demand equation.

238

Another popular choice as an instrumental variable is a policy change. States and municipalities frequently impose and change taxes and regulations that affect businesses. While these impositions can sometimes be a headache for businesses, they also can be useful when trying to measure causal relationships. For example, local sales taxes and/or price regulations can serve as instrumental variables for price in a demand equation. Or local labor laws (e.g., minimum wage) can serve as instrumental variables for wages when seeking to measure the effect of wages on productivity.

Broadly speaking, policy changes have strong potential as instrumental variables because they often affect business decisions (making them relevant) but often occur for reasons not related to business outcomes (making them exogenous). Typically, the ability to use a policy change as an instrumental variable will hinge on one's understanding of why the policy change occurred. If local sales tax rates are changing in response to retail demand

conditions, then they cannot be used as an instrumental variable for a demand equation whose outcome is retail sales. In contrast, if local sales tax rates are fluctuating due to budget issues driven by pensions for public employees, they likely are valid instruments, at least with regard to their exogeneity.

COMMUNICATING DATA 8.1

MEASURING THE IMPACT OF BROADBAND EXPANSION

Historically, a common point of interest among businesses and government is the impact of expanding broadband Internet access on economic development. In fact, in 2009 the U.S. government allocated \$7.2 billion toward broadband investment, at least in part with an expectation that it will have an economic impact. How can we tell whether broadband actually has such an impact?

The natural starting point is to collect data across regions in the U.S. on broadband availability along with economic measures (employment, productivity). We may express a proposed relationship (data-generating process) of the general form: $EconVar_i = \alpha + \beta BB_i + U_i$, where EconVar is an economic variable of interest and BB is broadband availability in the region. Within this formulation, it is highly likely that other factors influencing economic performance of a region would be correlated with broadband availability, precluding our ability to measure the effect of broadband availability on economic performance by running the corresponding regression. Adding controls may help this problem, but we may not have sufficient controls to solve it. Consequently, we may consider trying to find an instrumental variable for broadband availability. Such a variable must be relevant (correlated with broadband availability) and exogenous (not related to the economic variable of interest).

Following the intuition of the previous section, it can often be useful to consider variables that affect the cost of the treatment but do not affect the outcome per se. For our broadband problem, it has been proposed to use a measure of the slope of the local terrain as an instrumental variable, as steeper

landscapes can be more costly for broadband provision but likely not influential on economic outcomes per se. We could then perform 2SLS where we regress broadband availability on other controls along with a measure of slope in the first stage. Here, we can test whether slope is, in fact, relevant. The second stage uses the predicted values for broadband availability from the first stage and ultimately provides estimates and (properly adjusted for the fact that we are using predicted values) standard errors for the parameters of our assumed data-generating process. If the first stage shows slope to be relevant, and we are convinced via theoretical arguments that slope, per se, does not affect our economic variable of interest, 2SLS as described will provide us with a consistent estimate of the effect of broadband availability on an economic outcome.

239

Panel Data Methods

Panel data provide a unique opportunity for addressing endogeneity problems, rooted in the fact that, with panel data, we are able to observe the same cross-sectional unit (person, firm, etc.) multiple times at different points in time. The panel data methods we detail below are, in essence, an extension of the use of control variables to mitigate endogeneity. However, they warrant special mention due to the specific ways they are implemented and the interpretation of the associated estimates. We begin with the simplest application in the form of difference-in-difference regression, and then extend into more general applications using dummy variables and a within estimator.

DIFFERENCE-IN-DIFFERENCES

LO 8.4 Identify a difference-in-difference regression.

We begin our discussion of panel data methods by considering the case where we want to measure the effect of a dichotomous treatment. As an example, consider an individual who owns a large number of liquor stores in the states of Indiana and Michigan. Now suppose the Indiana state government decided to increase the sales tax on liquor sales by 3% beginning in January 2017. The store owner may want to learn the effect of this tax increase on her profits not only to “assess the damage” but also to form predictions about (and perhaps lobby against) possible future tax hikes in either state. To answer this question, the store owner may collect data on Profits each year for each store over a 2-year period spanning, say, 2016 and 2017. For each observation, she would also include whether that store experienced the tax hike imposed by the Indiana legislature. [Table 8.5](#) includes a hypothetical listing of the first few observations in such data.

To assess the effect of the tax hike on profits, the store owner may assume the following data-generating process:

$$\text{Profits}_{it} = \alpha + \beta \text{TaxHike}_{it} + U_{it}$$

Here, Profits_{it} is the profit of store i during Year t , and TaxHike_{it} equals 1 if the 3% tax hike was in place for store i during Year t and 0 otherwise. We could regress Profits on TaxHike, but it would be difficult to argue that TaxHike is not endogenous in such a regression. In particular, TaxHike equals 1 for a specific group of stores (in Indiana) at a specific time (2017), and this method of administering the treatment may be correlated with unobserved factors affecting Profits.

Why might observations receiving the treatment systematically have different profits (due to unobserved factors) than those not receiving the treatment? There are two clear

TABLE 8.5 Subsample of Profits for Liquor Stores in Indiana and Michigan in 2016 and 2017

STORE NUMBER	YEAR	STATE	PROFITS
1	2016	Indiana	\$65,000
1	2017	Indiana	81,000
2	2016	Michigan	47,000
2	2017	Michigan	32,000
3	2016	Indiana	35,000
3	2017	Indiana	51,000

240

reasons why this might be the case. First, all treated stores are in Indiana. People in Indiana may have different tastes for liquor, different income levels, different liquor regulations, etc., relative to Michigan; any of these factors could affect the profitability of a liquor store. For example, if people in Indiana generally have much stronger tastes for liquor than those in Michigan, we'd expect profits to be higher in Indiana whether those stores received the treatment or not. In this case, taste for liquor would be correlated with the tax hike, thus compromising our ability to measure the effect of the tax hike accurately.

The second reason treated observations may have different profits than untreated observations due to unobserved factors is that all treated observations were observed in 2017. Every time we observed a store that received the tax hike, we observed that store in 2017 (and not 2016). Liquor stores in 2017 may have different market conditions (unemployment rate), different liquor regulations, different competing products, etc., relative to liquor stores in 2016; any of these factors could affect the profitability of a liquor store at a given point in time. For example, if a federal law prohibiting marijuana is lifted in 2017, we might expect liquor-store profits to be lower in 2017 relative to 2016 if marijuana is a substitute for liquor. In this case, the lifting of the ban would be correlated with the tax hike, again compromising our ability to measure the effect of the tax hike accurately.

Fortunately, the panel nature of our data allows us to address endogeneity problems arising from unobservables that vary across the states and unobservables that vary across time. In essence, the solution is simply to add controls, but here the controls have a specific form. Specifically, the controls

mirror the structure of the panel data—we control for a cross-sectional group (g = Indiana, Michigan) and for time (t = 2016, 2017). In our tax example, this means we now assume the following data-generating process:

$$\text{Profits}_{igt} = \alpha + \beta_1 \text{Indiana}_g + \beta_2 \text{Year}_t + \beta_3 \text{TaxHike}_{gt} + U_{igt}$$

Here, i varies by store, g varies by state, and t varies by year. Further, Indiana_g equals 1 if the store is in Indiana and 0 otherwise, and Year_t equals 1 if the year is 2017 and 0 otherwise. In addition, TaxHike_{gt} equals 1 if both Indiana and Year equal 1 and 0 otherwise. Consequently, we can equivalently write the data-generating process as:

$$\text{Profits}_{igt} = \alpha + \beta_1 \text{Indiana}_g + \beta_2 \text{Year}_t + \beta_3 \text{Indiana}_g \times \text{Year}_t + U_{igt}$$

since $\text{Indiana} \times \text{Year}$ always equals TaxHike (they are both 1 when a store is in Indiana in 2017 and 0 otherwise).

With $\text{Indiana} \times \text{Year}$ indicating whether an observation received the treatment of the tax hike, the literal interpretation of β_3 is the effect of the tax hike, controlling for the state the store is in and the year in which it was observed. However, with these particular controls (for the cross-sectional unit and time), there is another way of interpreting β_3 . In particular, β_3 is the *difference-in-differences* for profits (or *diff-in-diff*, defined in general as follows) between years across states. Put another way, β_3 is the difference in the temporal change in profits between the treated state (Indiana) and the untreated state (Michigan). Restated once more, β_3 is the difference in how profits change over time between Indiana and Michigan.

We can use the assumed data-generating process from our example to help illustrate why β_3 is the diff-in-diff for profits in our example. For a given store, suppose we took

the difference in its profits between 2017 and 2016 for the case when it is in Indiana. This difference is:

$$\alpha + \beta_1 + \beta_2 + \beta_3 + U_{igt} - (\alpha + \beta_1 + U_{igt}) = \beta_2 + \beta_3$$

We can then make the same calculation for Michigan, comparing profits for a given store between 2017 and 2016:

$$\alpha + \beta_2 + U_{igt} - (\alpha + U_{igt}) = \beta_2$$

Lastly, we can take the difference between the change in profits in Indiana and the change in profits in Michigan to get the diff-in-diff:

$$\beta_2 + \beta_3 - \beta_2 = \beta_3$$

We illustrate this idea in [Figure 8.2](#).

Now that we've seen difference-in-differences in our specific example, let's generalize the concept. Consider a general case where we have two groups and two time periods—one group receives a treatment during the second time period. We assume the following data generating process for the outcome:

$$\text{Outcome}_{igt} = \alpha + \beta_1 \text{Treated}_g + \beta_2 \text{Period2}_t + \beta_3 \text{Treated}_g \times \text{Period2}_t$$

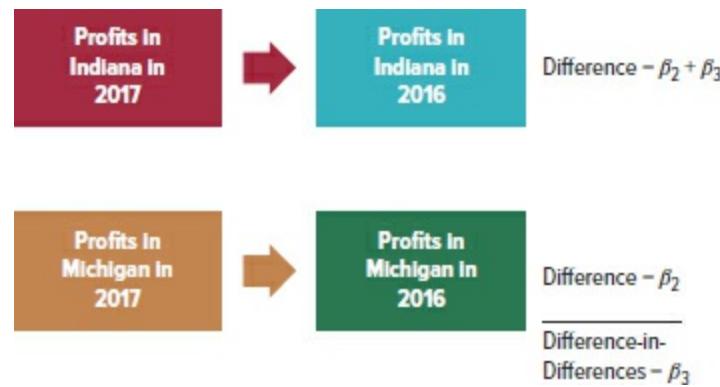
Here, Treated_g equals 1 if cross-sectional unit i is in the treated group (g) and

0 otherwise, and Period2_t equals 1 if the time period (t) is period 2 and 0 otherwise. Given this generalized data-generating process, the **difference-in-differences (diff-in-diff)** is defined as the difference in the temporal change for the outcome between the treated and untreated group. Again, the diff-in-diff is represented by β_3 in our data-generating process.

difference-in-differences (diff-in-diff) The difference in the temporal change for the outcome between the treated and untreated group.

By controlling for the group (treated vs. untreated) and the period (period 1 vs. period 2), we eliminate many possible confounding factors in the data-generating process. In particular, the control for the group controls for any factors affecting the outcome that differ across the groups but persist over time. For our tax example, we might believe, for example, that taste for liquor differs across the two states, but this difference is roughly persistent across 2016 and 2017. In contrast, the control for the period controls for any factors affecting the outcome that differ across time but are common across the groups. For our tax example,

FIGURE 8.2 Illustration of Difference-in-Differences for Liquor Profits in Indiana and Michigan



from 2016 to 2017, but these changes are largely common to both states.

While the controls for the group and the time period eliminate a great deal of possible confounding factors in our analysis, they do not eliminate all possible confounding factors. In our tax example, suppose there was a change to the legal drinking age in Indiana that took effect in January 2017. This policy change almost certainly would affect liquor profits and is clearly correlated with the treatment (the change in liquor taxes in Indiana). Hence, the change in drinking age is a confounding factor, and unfortunately, controlling for the group and time period will not eliminate it. The reason is that this confounding factor represents neither a persistent difference between Indiana and Michigan (and so the control for Indiana does not eliminate it), nor a difference across time that is common to both states (and so the control for the Year does not eliminate it). In [Reasoning Box 8.2](#) we summarize, in general, the circumstances in which diff-in-diff regression solves an endogeneity problem and those in which it does not.

LO 8.5 Execute regression incorporating fixed effects.

THE FIXED-EFFECTS MODEL

LO 8.6 Distinguish the dummy variable approach from a within estimator for a fixed-effects regression model.

The diff-in-diff model is highly effective, and applies for dichotomous treatments spanning two periods. However, many treatments in business and elsewhere are multi-level treatments and span multiple levels over several periods of time. The diff-in-diff model does

REASONING BOX 8.2

WHEN DOES DIFF-IN-DIFF
REGRESSION SOLVE AN
ENDOGENEITY PROBLEM?

Consider a population with two groups spanning two periods, where exactly one group receives a treatment in the second period. Assume the data-generating process: $\text{Outcome}_{igt} = \alpha + \beta \text{Treated}_g \times \text{Period2}_t + U_{igt}$, where i varies across N cross-sectional units, t varies between two periods, Treatment_g equals 1 if cross-sectional unit i is in the group (g) that received the treatment (and 0 otherwise), and Period2 equals 1 if $t = 2$ (and 0 otherwise). By construction, $\text{Treated} \times \text{Period2}$ equals 1 if the treatment was received and 0 otherwise.

IF:

Confounding factors in U are all either fixed over time or common among the treated and untreated groups

THEN:

Adding the controls Treated_g and Period2_t will eliminate endogeneity problems in measuring the effect of the treatment. In other words, we can assume the expanded data-generating process

$$\text{Outcome}_{igt} = \alpha + \beta_1 \text{Treated}_g + \beta_2 \text{Period2}_t + \beta_3 \text{Treated}_g \times \text{Period2}_t +$$

and running the corresponding diff-in-diff regression will yield a consistent estimate for β_3 .

In contrast, if confounding factors in U are changing over time in ways not common across the treated and untreated groups, then we cannot be assured the estimate for β_3 (the estimated effect of the treatment) is consistent.

USING DIFF-IN-DIFF TO ASSESS THE MINIMUM WAGE

A famous application of difference-in-differences was conducted by David Card and Alan Krueger in 1994. They attempted to measure the effect of a rise in the minimum wage on employment at fast-food restaurants. They used data on employment and the minimum wage in New Jersey (which raised its minimum wage from \$4.25 to \$5.05 in April 1992) and eastern Pennsylvania (which did not change its minimum wage around that time). Hence, all fast-food restaurants in New Jersey can be classified as “treated” stores, and all fast-food restaurants in eastern Pennsylvania can be classified as “untreated” stores. To simplify exposition, let's assume the researchers had data spanning the year before April 1992 and the year following April 1992. We could represent the data-generating process as:

$$\text{Employment}_{igt} = \alpha + \beta_1 \text{NewJersey}_g + \beta_2 \text{Year2}_t + \beta_3 \text{NewJersey}_g \times \text{Year2}_t$$

Here, g is either New Jersey or Pennsylvania, and t is either Year 1 (the year prior to the change) or Year 2 (the year after the change). Written this way, β_3 represents the difference-in-differences in employment between New Jersey and (eastern) Pennsylvania between Year 1 and Year 2.

We'd like to regress employment on an indicator for the state and an indicator for the year, along with their interaction, to get a consistent estimate for the effect of the minimum-wage increase on employment at fast-food restaurants. Put simply, if unobserved factors affecting employment are fixed over time or common across states, this regression will give us a consistent estimate. On the other hand, if there are factors changing over time in, say, New Jersey only, that affect employment, we cannot be confident that our estimate for β_3 really measures the effect of the minimum-wage change on employment. For example, if New Jersey also changes some of its regulations on fast-food restaurants (e.g., changes food-quality controls) during this time, we could not distinguish whether the change in employment (relative to Pennsylvania) was due to the minimum wage change or the regulations

change.

not directly apply to such cases, but it provides valuable intuition for a more general model that does—the fixed-effects regression model.

To illustrate, consider again our tax example, but now suppose we are interested in the effect of any change in the liquor tax on store profits—not just the effect of a single change (of 3%) in the tax rate. Further, suppose our store owner has stores across all states in the United States, and tax rates vary across the states and time. [Table 8.6](#) includes a hypothetical listing of the first few observations in such data. Here, we see tax rates vary across many levels (e.g., 3.5%, 6.2%, 7.5%), and the rates vary across the states and across time. Hence, there is not a dichotomous treatment (instead there are treatment levels), and there is no treatment period (treatment levels vary across multiple periods of time).

To assess the effect of a change in the liquor tax rate on profits, the store owner may assume the following data-generating process:

$$\text{Profits}_{it} = \alpha + \beta \text{TaxRate}_{it} + U_{it}$$

Here, Profits_{it} is again the profit of store i during Year t , and TaxRate_{it} is the liquor tax rate for store i during Year t . We could regress Profits on TaxRate, but it again would be difficult to argue that TaxRate is not endogenous in such a regression.

244

TABLE 8.6 Subsample of Tax Rate and Profits Data across States and Years

STORE NUMBER	YEAR	STATE	TAX RATE	PROFITS
1	2013	Indiana	3.2%	\$45,000
1	2014	Indiana	3.2%	31,000
1	2015	Indiana	4.1%	48,000

1	2016	Indiana	4.1%	52,000
1	2017	Indiana	5.0%	35,000
2	2013	Michigan	3.8%	51,000
2	2014	Michigan	4.5%	38,000
2	2015	Michigan	4.8%	41,000
2	2016	Michigan	5.0%	37,000
2	2017	Michigan	4.2%	43,000
3	2013	Ohio	6.1%	45,000
3	2014	Ohio	5.8%	49,000
3	2015	Ohio	6.1%	50,000

In the same vein as our diff-in-diff discussion, there are two clear reasons (but possibly more) why TaxRate might be correlated with unobservables affecting Profits. First, states likely differ in tastes for liquor, income levels, liquor regulations, etc., and any of these factors could affect the profitability of a liquor store. These differences also may affect the tax rates state legislators impose on liquor sales (high-income states may have higher taxes), generating an endogeneity problem. Second, there may be countrywide changes in market conditions, liquor regulations, competing products, etc., over time, and any of these factors could affect the profitability of liquor stores. These changes also may affect the tax rates state legislators impose on liquor sales (e.g., federal legalization of marijuana may generally result in lower liquor taxes), again generating an endogeneity problem.

As with our difference-in-differences version of the tax example, the panel nature of our data allows us to address endogeneity problems arising from unobservables that vary across the states and unobservables that vary across time. Again, the solution is simply to add controls. The controls we add when there are multiple states and multiple time periods are just an extension of the controls we added in the diff-in-diff model. Specifically, we add a dummy variable for each state (except one, which is the “base state”), and a dummy variable for each year (except one, which is the “base year”). In our tax example, this means we now assume the following data-generating process when there are G states and T years:

$$\text{Profits}_{igt} = \alpha + \delta_2 \text{State2}_g + \cdots + \delta_G \text{StateG}_g + \gamma_2 \text{Period2}_t + \cdots + \gamma_T]$$

For the above data-generating process, the literal interpretation of β is the effect of a change in the tax rate, controlling for the state the store is in and the year in which it was observed.

An alternative interpretation is not quite as simple as in the diff-in-diff model, but it is similar in spirit. Specifically, β measures the effect of a temporal change in the tax rate within a given state on store profits, relative to the overall trend in profits over that time period.

245

To elaborate, by controlling for states and time periods, β represents how profits for a given state move around its mean, relative to the overall trend of states' profits, when there is a change in the tax rate. As an example, suppose mean profits over the T years in New York were \$100,000, and the overall trend in profits was an increase of \$1,000 per period. Then, if T were 5, we'd expect profits in New York, absent any change in the tax rate, to look like: \$98,000 in Year 1, \$99,000 in Year 2, ..., and \$102,000 in Year 5. Here, β measures how much New York profits would deviate from this trend when the tax rate changes—for example, how much different profits would be compared to \$99,000 for a change in the tax rate in Year 2.

Let's now generalize the concept of our extended tax example. Consider a general case where we have G groups and T time periods, and a treatment (dichotomous or multi-level) that varies across groups and across time. We assume the following data-generating process for the outcome:

$$\begin{aligned} \text{Outcome}_{igt} = & \alpha + \delta_2 \text{Group2}_g + \cdots + \delta_G \text{GroupG}_g + \gamma_2 \text{Period2}_t \\ & + \cdots + \gamma_T \text{PeriodT}_t + \beta \text{Treatment}_{gt} + U_{igt} \end{aligned}$$

Here, we have dummy variables for each group (except for the base group,

generically chosen to be Group1) and dummy variables for each time period (except for the base period, generically chosen to be Period1). Outcome_{*igt*} is the observed outcome for observation *i* in group *g* at time *t*, and Treatment_{*gt*} is the level of treatment experienced by cross-sectional group *g* in period *t*. The above generalized data-generating process is an example of a **fixed effects model**, defined as a data-generating process for panel data that includes controls for cross-sectional groups. The controls for cross-sectional groups are called **fixed effects**.

fixed effects model A data-generating process for panel data that includes controls for cross-sectional groups.

fixed effects The controls for cross-sectional groups.

We conclude this section by making several important points concerning the general fixed effects model. First, for a data-generating process to be characterized as a fixed effects model, it need have only controls for the cross-sectional groups; it need not have controls for the time periods. However, we present the model with controls for time periods, as their inclusion is generally advisable in practice. If controls for time are excluded, this runs the risk of, for example, the treatment and outcome trending over time for all cross-sectional groups, and mistaking this co-movement as the effect of the treatment. In our tax example, this type of problem would manifest if tax rates and profits were both trending up over time everywhere (likely for differing reasons). Without controls for time, it will appear as though profits were increasing over time because of the increase in tax rates, when it was instead trending market-level factors (generally worsening state budgets and enhanced tastes for liquor for the taxes and profits, respectively) that were behind each trend separately.

A second point concerning our general fixed effects model concerns the structure of the time controls. In the model we present, there is a separate dummy variable for each time period. This is the most flexible way to control for time, as it puts no restrictions on how profits change period-to-period for all the cross-sectional units. However, it is common to instead control for

time periods by including a time trend. By doing so, the data-generating process is:

$$\text{Outcome}_{igt} = \alpha + \delta_2 \text{Group2}_g + \cdots + \delta_G \text{GroupG}_g + \gamma \text{Time}_t + \beta \text{Treat}$$

Here, Time_t is equal to the time period (e.g., it equals one in period 1, two in period 2, and so on). This alternative formulation forces the effect on profits of moving from one

246

period to the next to be constant (equal to γ). This restriction is sensible in many applications, where variables are likely to be generally trending upward or downward over time, and requires us to estimate fewer parameters. In contrast, it is inappropriate to make this simplification when the outcome is trending in uneven ways (e.g., up and then down, or up but by a notably differing rate).

Thirdly, as in the diff-in-diff model, by controlling for the groups (adding fixed effects) and the periods, we eliminate many possible confounding factors in the data-generating process. Again, the controls for the groups control for any factors affecting the outcome that differ across the groups but persist over time. And the controls for the periods control for any factors affecting the outcome that differ across time but are common across the groups. However, as in the diff-in-diff model, these controls do not eliminate all possible confounding factors. In particular, these controls do not eliminate endogeneity problems arising from unobserved factors affecting the outcome that change over time for only a subset of groups in a way that is correlated with the treatment. In our tax example, suppose New York and New Jersey uniquely experienced an increase in residents' tastes for liquor over time concurrent with declines in their liquor tax rates. Given the change in tastes would likely generate higher profits for liquor stores, we would be unable to determine whether any changes we observe in store profits are due to the change in tastes or the lowering of taxes, despite our controls.

Lastly, we note that we can always add controls (X_{igt} 's) beyond the fixed effects and time dummies to help eliminate some of the remaining confounding factors. Such controls are able to help with endogeneity problems only if they vary over time in different ways for the different cross-sectional groups; otherwise, they are collinear with the fixed effects or the time dummies and therefore must be dropped from the model. To illustrate, in our tax example, we may have data on the number of competing liquor stores in each state; however, if this variable is fixed over time, it will be perfectly collinear with our fixed effects and thus must be dropped from the model. In contrast, if the number of competing liquor stores not only differs across states but also evolves over time differently across states, adding a variable (e.g., Competitors_{gt}) could help mitigate remaining endogeneity concerns, as it may eliminate a confounding factor that varies across states and years.

Now that we have defined and described the fixed-effects model, we detail two alternative ways of estimating it.

Dummy Variable Estimation Recall our formulation of a general fixed-effects model as:

$$\begin{aligned}\text{Outcome}_{igt} = & \alpha + \delta_2 \text{Group2}_g + \cdots + \delta_G \text{GroupG}_g + \gamma_2 \text{Period2}_t \\ & + \cdots + \gamma_T \text{PeriodT}_t + \beta \text{Treatment}_{gt} + U_{igt}\end{aligned}$$

The first method we consider for estimating this model is known as the dummy variable method, or dummy variable estimation, for a fixed effects model. **Dummy variable estimation** uses regression analysis to estimate *all* of the parameters in the fixed effects data-generating process. Dummy variable estimation involves performing regression analysis exactly as the data-generating process would suggest—regress the Outcome on dummy variables for each cross-sectional group (except the base unit), dummy variables for each period (except the base period), and the treatment.

dummy variable estimation Uses regression analysis to estimate *all* of the parameters in the fixed effects data-generating process.

247

TABLE 8.7 Subset of Dummy Variable Estimation Results for Sales Regressed on TaxRate

	COEFFICIENTS	STANDARD ERROR	t STAT
Intercept	38147.38615	1079.619121	35.33411497
State2	5327.21415	712.7437924	7.474234369
State3	1903.641928	721.1620734	2.639686692
State4	-3703.480692	749.339573	-4.942326317
Year2	1158.939473	750.0000243	1.54525258
Year3	2284.199776	734.0625504	3.111723619
Year4	3162.033836	721.4070438	4.383147992
TaxRate	-925.6772431	170.4044168	-5.432237383

In Table 8.7 we present a subset of regression results for our generalized tax example. Here, we have parameter estimates for each state (except State 1), each year (except Year 1), and the Tax Rate. Proper interpretation of these results is as important as proper modeling, so let's carefully interpret each type of coefficient (state, year, tax rate) in turn.

We start with the state coefficients. Each state coefficient measures the effect on a store's profits of moving the store from the base state (State 1) to that alternative state, for a given year and tax rate. For example, the coefficient on State 2 implies that moving a store from State 1 to State 2 will increase profits by about \$5,327, holding the year and tax rate constant.

Interpretation is similar for the year coefficients. Each year coefficient measures the effect on a store's profits of moving the store from the base year (Year 1) to that alternative year, for a given state and tax rate. For example, the coefficient on Year 2 implies that moving a store from Year 1 to Year 2 will increase profits by about \$1,159, holding the state and tax rate constant.

Lastly, the coefficient on Tax Rate measures the effect on a store's profits

of changing the Tax Rate, for a given state and year. The coefficient on Tax Rate in [Table 8.7](#) implies that increasing the Tax Rate by 1 percentage point will decrease profits by about \$926, holding the state and year constant.

Returning to our general fixed effects model:

$$\text{Outcome}_{igt} = \alpha + \delta_2 \text{Group2}_g + \cdots + \delta_G \text{GroupG}_g + \gamma_2 \text{Period2}_t + \cdots \\ + \gamma_T \text{PeriodT}_t + \beta \text{Treatment}_{gt} + U_{igt}$$

we can again detail how to interpret each type of parameter estimate (group, period, treatment). An estimate for δ_j measures the effect of moving from Group 1 to Group j for a given Period and Treatment level. Note also that we can measure the effect of moving from Group j to Group k (holding Period and Treatment constant) by taking the difference in their estimated coefficients: $\hat{\delta}_K - \hat{\delta}_j$. Similarly, an estimate for γ_q measures the effect of moving from Period 1 to Period q for a given Group and Treatment level. Lastly, β measures the effect of the Treatment on the Outcome for a given Group and given Period. Notice how this last interpretation clearly highlights the type of confounding factors that cannot bias the estimate—factors that are fixed across periods for the groups or fixed across groups for a period do not pose endogeneity problems, as the model controls for them.

248

8.3

Demonstration Problem

Suppose you have data on an outcome variable, Y , that varies by month and by employee for a fixed (over time) group of employees. Each employee works in one of six designated regions in the United States, recorded as the variable “Region” and taking on the values 1 through 6. The data span 8 months; they also contain information on a treatment, which ranges from 0 to 10 and varies

across regions and months.

- a. Write out the fixed effects model you would assume in trying to determine the effect of the treatment on the outcome.
- b. Interpret the coefficient on the treatment within your fixed effects model.
- c. Explain the role of the fixed effects within your fixed effects model.

Answer:

- a. $Y_{igt} = \alpha + \delta_2 \text{Region2}_g + \dots + \delta_6 \text{Region6}_g + \gamma_2 \text{Month2}_t + \dots + \gamma_8 \text{Month8}_t$. Here, i varies by employee, g varies by region (1 to 6), and t varies by month (1 to 8).
- b. The coefficient on the treatment (β) represents the effect of a change in the treatment on the outcome for a given region during a given month. For example, if an employee experiences an increase in the treatment by one during, say, month 3, and stays in, say, region 5, her outcome will change by β .
- c. The fixed effects (Region2 ... Region6) control for time-invariant factors affecting the outcome within each Region.

Within Estimation A notable issue with using dummy variable estimation to estimate a fixed effects model is that it may require us to estimate a very large number of parameters. For example, suppose we have data on individuals over time and want to include fixed effects for each individual. If we have 100,000 different individuals, we must then estimate the coefficients on 99,999 fixed effects. This can be time-consuming, even for fast computers. Further, it is often the case that we are not interested in the effect of changing groups per se, but care only about the effect of the treatment. To illustrate, in our tax example, we likely do not care about the effect on profits of moving a store from one state to another if the analysis is focused on measuring the impact of the tax. The purpose is not to assess where to open or move stores but rather to evaluate the tax. In such a circumstance, we'd like to estimate the effect of the tax, controlling for groups and time periods,

but without having to estimate the effects of what may be a very large number of groups.

Fortunately, there is a simple alternative method of estimating a fixed effects model that eliminates the need to estimate the coefficient for each fixed effect. **Within estimation** uses regression analysis of within-group differences in variables to estimate the parameters in the fixed effects data-generating process, except for those corresponding to the fixed effects (and the constant). Perhaps the best way to understand exactly how and why within estimation works is to see it in practice.

within estimation Uses regression analysis of within-group differences in variables to estimate the parameters in the fixed effects data-generating process, except for those corresponding to the fixed effects (and the constant).

Consider again our expanded tax example, where we've assumed the following data-generating process:

$$\text{Profits}_{igt} = \alpha + \delta_2 \text{State2}_g + \cdots + \delta \text{State}G_g + \gamma_2 \text{Year2}_t + \cdots + \gamma_T \text{Year}T_t + \epsilon_{igt}$$

249

If we have data spanning all 50 states, we have 49 δ 's to estimate for the fixed effects if we use dummy variable estimation. Consider now the following alternative: For each state (group), we add up the data-generating process for each store and each year, and then divide by $N_g \times T$ to get the state average, where N_g is the number of stores in state g . For a state (group) g , this calculation looks as follows:

$$\frac{1}{N_g T} \sum_{i=1}^{N_g} \sum_{t=1}^T \text{Profits}_{igt} = \overline{\text{Profits}}_g$$

Equivalently:

$$\begin{aligned} \frac{1}{N_g T} \sum_{i=1}^{N_g} \sum_{t=1}^T & \alpha + \delta_2 \text{State2}_g + \cdots + \delta_G \text{StateG}_g + \gamma_2 \text{Year2}_t + \cdots \\ & = \alpha + \delta_2 \text{State2}_g + \cdots + \delta_G \text{StateG}_g + \gamma_2 \left(\frac{1}{T} \right) + \cdots + \gamma_T \left(\frac{1}{T} \right) + \end{aligned}$$

Notice that the group-level average for the State dummies is equal to their values for each observation in the group; this is because they are constant within a group. For example, all observations in State 2 have $\text{State2} = 1$ for each store and each year; thus, its average is 1. Also, notice that the year dummies always average to $1/T$, since each equals 1 in exactly one period and 0 for all others within a group.

Next, we take each observation in our data and subtract its group-level average. This calculation looks as follows:

$$\begin{aligned} \text{Profits}_{igt} &= \alpha + \delta_2 \text{State2}_g + \cdots + \delta_G \text{StateG}_g + \gamma_2 \text{Year2}_t \\ &\quad + \cdots + \gamma_T \text{YearT}_t + \beta \text{TaxRate}_{gt} + U_{igt} \\ - \overline{\text{Profits}}_g &= \alpha + \delta_2 \text{State2}_g + \cdots + \delta_G \text{StateG}_g + \gamma_2 \left(\frac{1}{T} \right) \\ &\quad + \cdots + \gamma_T \left(\frac{1}{T} \right) + \beta \overline{\text{TaxRate}}_g + \overline{U}_g \\ (\text{Profits}_{igt} - \overline{\text{Profits}}_g) &= \left(\frac{1}{T} \right) (\gamma_2 + \cdots + \gamma_T) + \gamma_2 \text{Year2}_t + \cdots + \gamma_T \\ &\quad + \beta (\text{TaxRate}_{gt} - \overline{\text{TaxRate}}_g) + (U_{igt} - \overline{U}_g) \end{aligned}$$

Let's now define a few new variables as follows:

$$\begin{aligned} \text{Profits}_{igt}^* &= \text{Profits}_{igt} - \overline{\text{Profits}}_g, \\ \text{TaxRate}_{gt}^* &= \text{TaxRate}_{gt} - \overline{\text{TaxRate}}_g \end{aligned}$$

and

$$U_{igt}^* = U_{igt} - \bar{U}_g$$

Further, define

$$\alpha^* = \left(\frac{1}{T} \right) (\gamma_2 + \cdots + \gamma_T)$$

With these definitions and the calculations above, we have identified a derivative data-generating process involving these new variables as follows:

$$\text{Profits}_{igt}^* = \alpha^* + \gamma_2 \text{Year2}_t + \cdots + \gamma_T \text{YearT}_t + \beta \text{TaxRate}_{gt}^* + U_{igt}^*$$

250

This newly defined data-generating process expresses profits relative to the group mean as depending on: the year, the Tax Rate relative to its group mean, and unobservables relative to their group mean. This formulation focuses on relative performance of the treatment and outcome within each group; consequently, it ignores any relationship between tax rates and profits due to differences across groups. By ignoring cross-group differences in the treatment and outcome, we eliminate the possibility that the relationship we find between the treatment and outcome is due to differences in cross-group unobservables (e.g., market conditions) rather than a causal effect of the tax rate on profits.

Notice that, with our new formulation, we retain nearly all of the same parameters as our original formulation; $\gamma_2, \dots, \gamma_T, \beta$ are all the same. Therefore, attaining a consistent estimate of β (the causal effect of the

treatment) by regressing Profits* on TaxRate* and the Year dummies essentially depends on whether we believe there exists an endogeneity problem within this newly defined data-generating process. In particular, we must ask whether U^* is uncorrelated with TaxRate*. In words, we must believe that differences in unobservables (e.g., state-level market conditions) from their state-level means (across stores and years in that state) are not correlated with differences in Tax Rates from their state-level means. This assumption would fail, for example, if we believed state-level tax rates were dependent on state-level unemployment and this latter variable affected state-level liquor store profits.

We conclude by laying out the within estimation steps for a general fixed effects model. Consider a general fixed effects model:

$$\begin{aligned} \text{Outcome}_{igt} = & \alpha + \delta_2 \text{Group2}_g + \cdots + \delta_G \text{GroupG}_g + \gamma_2 \text{Period2}_t \\ & + \cdots + \gamma_T \text{PeriodT}_t + \beta \text{Treatment}_{gt} + U_{igt} \end{aligned}$$

We estimate the parameters $\gamma_2, \dots, \gamma_T, \beta$ via within estimation by executing the following steps:

1. Determine the cross-sectional groups (e.g., states) and calculate group-level means: $\overline{\text{Outcome}}_g = \frac{1}{N_g T} \sum_{i=1}^{N_g} \sum_{t=1}^T \text{Outcome}_{igt}$ and $\overline{\text{Treatment}}_g = \frac{1}{N_g T} \sum_{i=1}^{N_g} \sum_{t=1}^T \text{Treatment}_{igt}$.
2. Create new variables:
 $\text{Outcome}_{igt}^* = \text{Outcome}_{igt} - \overline{\text{Outcome}}_g$, $\text{Treatment}_{igt}^* = \text{Treatment}_{igt} - \overline{\text{Treatment}}_g$. (Note: if the assumed model includes controls (X s), for each X , create $X_{igt}^* = X_{igt} - \overline{X}_g$.)
3. Regress Outcome^* on Treatment^* and the Period dummy variables.
(Again, if the assumed model includes controls (X s), include X^* 's in the regression.)

Comparing Estimation Methods For a fixed effects model, we can utilize either dummy variable estimation or within estimation to estimate the parameters of the model. In this section, we briefly compare these two options. First, note that both methods yield exactly the same estimates for the effect of the treatment (β) and time periods ($\gamma_2, \dots, \gamma_T$). The two estimation methods differ in their estimates for the constant, but this is typically inconsequential. Next, dummy variable estimation provides estimates for the fixed effects (the effect of switching groups on the outcome), whereas within estimation does not. If the effect of group switching is of interest, then dummy variable estimation is clearly preferred. If not, within estimation is likely preferred as it provides an estimate for the effect of the treatment without wasting computation time on the fixed effects.

251

8.4 Demonstration Problem

As in [Demonstration Problem 8.3](#), suppose you have data on an outcome variable, Y , that varies by month and by employee for a fixed (over time) group of employees. Each employee works in one of six designated regions in the United States, recorded as the variable “Region” and taking on the values 1 through 6. The data span 8 months and also contain information on a treatment, which ranges from 0 to 10 and varies across regions and months.

Write out the data-generating process suitable for within estimation in trying to determine the effect of the treatment on the outcome. Be sure to define each variable in the expression you compose.

Answer:

The data-generating process suitable for within estimation is:

$$Y_{igt} - \bar{Y} = \left(\frac{1}{8}\right)(\gamma_2 + \dots + \gamma_8) + \gamma_2 \text{Month2}_t + \dots + \gamma_8 \text{Month8}_t \\ + \beta (\text{Treatment}_{gt} - \overline{\text{Treatment}}_g) + (U_{igt} - \bar{U}_g)$$

Here, the dependent variable is the difference in the outcome of employee i in region g during month t from the average for that region across all employees in that region and all 8 months.

The independent variables consist of dummy variables for each month, and the difference in the treatment for region g during month t (which is the same for all employees in that region during that month) from the average treatment for that region across all employees in that region and all 8 months.

Lastly, the new term for the unobservables is the difference in the unobservables for employee i in region g during month t from the average unobservables for that region across all employees in that region and all 8 months.

Another point of comparison between these two estimation methods pertains to R -squared. As noted in [Chapter 6](#), we should not place much weight on R -squared in assessing a model of causality that is suitable for active predictions. However, even in this context, it is sometimes used to get a sense of how strongly the model fits the data. For dummy variable estimation, the R -squared is often misleadingly high, suggesting a very strong fit. This is because, for dummy variable estimation, we are trying to explain variation in the Outcome, and much of this variation may be explained by the fixed effects (differences across groups). Hence, the R -squared may be high because of inclusion of fixed effects and not because variation in the Treatment explains much of the variation in the Outcome. In contrast, for within estimation, we are trying to explain variation in the Outcome relative to its group-level mean. Here, the fixed effects are dropped, so a high R -squared is more indicative that variation in the Treatment (relative to its

mean) is explaining variation in the Outcome (relative to its mean).

Lastly, we note that whether we use dummy variable estimation or within estimation to estimate the parameters of a given fixed effects model, we are susceptible to the same concerns about endogeneity. Specifically, both estimation methods eliminate confounding factors that are fixed across periods for the groups or are fixed across groups over time.

252

REASONING BOX 8.3

IMPLICATIONS OF THE FIXED EFFECTS MODEL

IF:

We assume the following data-generating process:

$$\begin{aligned}\text{Outcome}_{igt} = & \alpha + \delta_2 \text{Group2}_g + \dots + \delta_G \text{GroupG}_g + \gamma_2 \text{Period2}_t \\ & + \dots + \gamma_T \text{PeriodT}_t + \beta \text{Treatment}_{gt} + U_{igt}\end{aligned}$$

THEN:

1. β is the effect of the Treatment on the Outcome for a given Group during a given Period.
2. No factors in U that are fixed across periods for the groups or are fixed across groups over time can generate an endogeneity problem in this model.
3. Estimation using dummy variable estimation or within estimation yields the same estimate for β (and $\gamma_2, \dots, \gamma_T$).

And both estimation methods could yield inaccurate (inconsistent) estimates if there exist unobserved factors that vary within a group over time.

We conclude by summarizing some basic implications of the fixed effects

model that can aid in interpretation regardless of the estimation method as shown in [Reasoning Box 8.3](#).

PRACTICAL APPLICATIONS OF PANEL DATA METHODS FOR BUSINESS

In this section, we discuss practical applications of panel data methods for business along two dimensions. First, we highlight the types of panel data one is likely to encounter in business settings. We then discuss the process of grouping these data, which ultimately determines the fixed effects to include in the assumed data-generating process.

For a given firm, panel data typically have cross-sectional units that are either individuals/households or divisions/branches. A firm may conduct or purchase a survey that questions a set of individuals at multiple points in time. The survey may ask about purchases the individual made or preferences for different products. Using such data, analysts can apply panel data methods to learn about things like the effect of advertising on preferences or the effect of new product introductions on demand for existing products. As another example, a firm may collect information on production and defects at multiple points in time from multiple production facilities. These panel data could be used to learn about the effect of new maintenance programs on the incidence of defective products. Lastly, firms are able to collect a wealth of panel data via their websites by asking customers to register and log in whenever they make a purchase. This allows for the use of panel data methods, as they observe many individuals on many separate occasions.

In our tax example, the panel data were such that the cross-sectional unit was a store and the time period was a year. We chose to group these data by state rather than store. However, we could have grouped these data by store instead. What dictates the decision of how to form groups (and thus construct the corresponding fixed effects) in a fixed

the same tax rate. Consequently, unobservables affecting profits that vary across stores within a given state and year cannot create an endogeneity problem for the tax rate, since they cannot be correlated with the tax rate (the tax rate is the same for all of them). The upshot from this example is that there is no value in adding fixed effects that vary at a finer level in the cross-section than the treatment (i.e., stores vs. states).

Following the intuition of our tax example, we now make two general points on the grouping process for fixed effects models:

- First, when building a fixed effects model, the groupings should be no finer than the cross-sectional groups along which the treatment varies. Grouping at too fine a level places more demands on the data (more parameters to estimate or less variation in the redefined variables) without helping mitigate an endogeneity problem.
- Second, there are occasions when it might be optimal or necessary to choose groupings that are coarser than the cross-sectional groups along which the treatment varies. It may be optimal to choose coarser groupings if you are convinced that confounding factors vary only across coarse cross-sectional groups. In our tax example, we could have created groups according to geographic region (Northeast, Southwest, etc.) rather than states if we believed confounding factors varied only at that level. It may be necessary to choose coarser groupings if the data are relatively small. Again, in our tax example, we may need to group according to geographic region rather than states if, say, we had only 150 observations. In such a circumstance, there would likely be too little information left in the data after controlling for each state to measure the effect of the tax rate, and so fewer fixed effects must be used. Carrying through with the analysis using coarser groupings (and thus fewer fixed effects) must be weighed against the risk of having an inconsistent estimate if regional fixed effects are not sufficient to deal with endogeneity problems.

COMMUNICATING DATA 8.3

DOES MULTIMARKET CONTACT AFFECT AIRLINES' ON-TIME PERFORMANCE?

A pervasive question relevant to business strategy is whether contact across many markets facilitates firms' ability to soften competition. Along with Daniel Simon, I have investigated this question for airlines, seeking to establish whether contact across many different routes leads to poorer on-time performance (suggestive of softer competition) by airlines. To tackle this question, we collected data on travel time and multimarket contact for millions of flights spanning many airlines, routes, and quarters. Here, multimarket contact measures the number of times the carrier for a given flight interacts with other airlines serving the same route during that quarter across other routes. A simplified version of our assumed data-generating process is:

$$\begin{aligned}\text{TravelTime}_{irt} = & \alpha + \delta_2 \text{CarrRoute2}_{ir} + \dots + \delta_{IR} \text{CarrRouteIR}_{IR} \\ & + \gamma_2 \text{Quarter2}_t + \dots + \gamma_T \text{QuarterT}_t + \beta \text{MMC}_{irt} + U_{irt}\end{aligned}$$

In this formulation, we created cross-sectional groups of carrier-routes, where there are I carriers and R routes. For example, one group consists of all flights by United between Chicago and Atlanta; thus, the carrier-route is United on

254

Chicago-Atlanta. The variable MMC_{irt} is a variable that measures the amount of times carrier i interacts with the other carriers on route r during quarter t on other routes.

We regress travel time on our fixed effects, quarter dummy variables, and multimarket contact. This formulation ensures that our estimate for the effect of multimarket contact is immune from endogeneity issues arising from time-invariant factors affecting travel time for each carrier-route combination and

from general trends in travel time over time. For example, suppose travel time for United on the Chicago-Atlanta route tends to be long due to congestion problems, but these problems persist over time. Then, our fixed effect for United on the Chicago-Atlanta route will control for these congestion factors, thus eliminating them as possible confounding factors in the regression.

Interestingly, our analysis showed that multimarket contact did in fact worsen on-time performance, and the strength of this conclusion is bolstered by the fact that our fixed-effects analysis removed many potential confounding factors. This finding suggests softer competition—at least on on-time performance—when firms come in contact across many markets.

RISING TO THE dataCHALLENGE

Do TV Ads Generate Web Traffic?

The goal for this challenge is to measure the effect on website visits of running an ad in a county. We may want to know the immediate effect of the ad—that is, the change in website visits on the day the ad is run. We have data on counties on a daily basis, meaning we have panel data at the county-day level. The simplest model we may consider to ultimately measure the effect would be:

$$\text{Visits}_{it} = \alpha + \beta Ad_{it} + U_{it}$$

Here, Visits_{it} is the number of visits to the website in county i on day t and Ad_{it} equals 1 if the ad was run in county i on day t and 0 otherwise. Given the ads were deliberately targeted to specific counties, we should be worried that there are characteristics of the counties receiving the ad that also influence the number of visits to the [FunnyHa.com](#) website.

Since we have panel data, we can control for time-invariant county-level differences using county fixed effects, and we can control for factors varying over time that are common to all counties. After we include these controls, our

assumed data-generating process becomes:

$$\begin{aligned} \text{Visits}_{it} = & \alpha + \delta_{\text{County2}} 2_i + \cdots + \delta_G \text{CountyG}_i \\ & + \gamma_2 \text{Day2}_t + \cdots + \gamma_T \text{DayT}_t + \beta Ad_{it} + U_{it} \end{aligned}$$

We can estimate this model using dummy variable estimation or within estimation. Either way, the estimate for β will be the same. It can be interpreted as the effect on website visits of running the ad that day for a given county on a given day.

255

Note that an endogeneity problem may still exist. This will be the case if there are unobserved factors influencing website visits that vary differently in the counties over time in a way that is correlated with the timing of the ad being run in the county. If this concern is substantial enough, you may want to attempt to find an instrumental variable for Ad (a variable that is correlated with where and when an ad was run, but not related to the number of visits to the website).

SUMMARY

In this chapter we presented two relatively advanced methods for establishing causal inference—use of instrumental variables and panel data methods. We defined instrumental variables and illustrated how they can aid in measuring causal effects. We went on to explain how to utilize instrumental variables via two-stage least squares regression, and then discussed how to evaluate the validity of instruments in terms of exogeneity and relevance. We concluded our discussion of instrumental variables by highlighting some of their classic practical applications.

We opened our discussion of panel data methods by detailing difference-in-differences and how it applies to circumstances where there is a dichotomous treatment and just two time periods. We built on these ideas as we presented the fixed effects model and illustrated how fixed effects (and time controls) can help mitigate endogeneity problems. Next, we showed how to estimate a fixed

effects model using dummy variable estimation and within estimation, and compared and contrasted the two approaches. We concluded by highlighting some practical business applications of panel data methods.

KEY TERMS AND CONCEPTS

difference-in-differences (diff-in-diff)

dummy variable estimation

exogeneity of an instrumental variable

fixed effects

fixed effects model

instrumental variable

relevant as an instrumental variable

two stage least squares (2sls) regression

weak instrument

within estimation

CONCEPTUAL QUESTIONS connect

1. Suppose you have assumed the following data-generating process:
-

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + U_i$$

Suppose also that you believe X_2 is endogenous in this model. Which of the following variables would be valid instruments for X_2 ? (LO3)

- Z_1 , where $\text{spCorr}(Z_1, X_1(X_2, X_3)) = 0$ & $\text{Corr}(Z_1, U) = 0$
- Z_2 , where $\text{spCorr}(Z_2, X_2(X_1, X_3)) \neq 0$ & $\text{Corr}(Z_2, U) \neq 0$
- Z_3 , where $\text{spCorr}(Z_3, X_2(X_1, X_3)) \neq 0$ & $\text{Corr}(Z_3, U) = 0$
- Z_4 , where $\text{spCorr}(Z_4, X_3(X_1, X_2)) \neq 0$ & $\text{spCorr}(Z_4, X_1(X_2, X_3)) \neq 0$

2. Refer to Question 1. Suppose you have a valid instrument, Z , for X_2 . (LO2)
 - a. What are the steps that would be taken as part of the first stage of 2SLS estimation?
 - b. What are the steps that would be taken as part of the second stage of 2SLS estimation?
3. Again, refer to Question 1. Suppose you have a candidate instrument, Z , for X_2 . (LO3)
 - a. Explain how you can test whether Z is exogenous in this model.
 - b. Explain how you can test whether Z is relevant in this model.
4. Suppose you are interested in learning the effect of an increase in commission rates on sales for your workforce. On July 1 of last year, your East Coast managers implemented an increase in commission rates, while your West Coast managers chose not to give the increase. You have sales data for all of last year for each employee on each coast. (LO4)
 - a. We can use difference-in-differences to measure the effect of the increase in commission rate on sales using these data. What is the difference-in-differences for this problem?
 - b. Write out a data-generating process that allows for a measurement of difference-in-differences.
 - c. What are some possible confounding factors within the data-generating process you constructed for Part b?
5. Assume the following data-generating process:

$$Y_{igt} = \alpha + \delta_2 \text{Group2}_g + \delta_3 \text{Group3}_g + \delta_4 \text{Group4}_g + \gamma_2 \text{Period2}_t + \gamma_3$$
 - a. Interpret δ_3 .
 - b. Interpret γ_4 .
 - c. Interpret β .
 - d. What is the effect of moving from Group 2 to Group 4, holding all other factors constant?
6. Refer to Question 5. Construct a new data-generating process that is suitable for within estimation. (LO6)
7. Suppose you are interested in the causal effect of price on your product's sales. However, you recognize that simply regressing sales on price will

almost certainly suffer from an endogeneity problem. Your manager asks how to deal with this problem, and you suggest finding an instrumental variable, Z . Explain in nontechnical terms: (LO1)

- a. The properties Z must possess in order to be a valid instrument for this particular problem.
 - b. How, if Z is a valid instrument, it can allow us to measure the causal effect of price on sales.
8. You are interested in estimating the effect of X on Y in the following assumed data-generating process: $Y_i = \alpha + \beta X_i + U_i$. You are concerned that there is an endogeneity problem if you simply regressed Y on X and are considering using a variable, Z , as an instrument. Explain in nontechnical terms why Z cannot help in establishing causality for X if Z is: (LO1)
- a. Not relevant.
 - b. Not exogenous.
9. Suppose you have panel data that span two time periods and two groups, where Group 2 received a treatment in Period 2 and Group 1 never received a treatment. Suppose also that you assume the following data-generating process for
- $$Y : Y_{igt} = \alpha + \delta \text{Group2}_g + \gamma \text{Period2}_t + \beta \text{Group2}_g \text{Period2}_t + U_{igt}.$$
- Here, Group2_g equals 1 if the observation is from Group 2 and 0 otherwise, and Period2_t equals 1 if the observation is from Period 2 and 0 otherwise. (LO4)
- a. Interpret β .
 - b. Explain carefully the potential consequence of excluding the variable Group2 when running a regression designed to measure the average treatment effect.
 - c. Explain carefully the potential consequence of excluding the variable Period2 when running a regression designed to measure the average treatment effect.

QUANTITATIVE PROBLEMS connect

- 10.** You are working as an analyst for a large cable company that offers bundles of channels all across the United States. One of the bundles is the “basic package,” which includes network channels along with a few other basic cable channels. You are interested in learning how the price of this basic package influences the rate of subscriptions in a market. You have data on subscriptions per 1,000 local residents, price for the basic package, and average local household income. You also have data on local telecom labor costs per subscriber. You believe this last variable influences the local price but not subscriptions per se.

Dataset available at www.mhhe.com/prince1e

Use the data in *Chap8Prob10.xlsx* for this question. (LO2)

- a. Based on the data provided, write out an expression for the data-generating process for subscriptions per 1,000 local residents.
- b. Estimate the effect of basic package price on subscriptions per 1,000 local residents using OLS. Why might you distrust this result as being a causal effect?
- c. Estimate the effect of basic package price on subscriptions per 1,000 local residents using 2SLS.
- d. Using your 2SLS results, what is a 99% confidence interval for the effect of the basic package price on subscriptions per 1,000 local residents?

Dataset available at www.mhhe.com/prince1e

- 11.** Refer to Problem 10. Use the data in *Chap8Prob11.xlsx* for this question. (LO2)
- a. As you did in Part c of Problem 10, estimate the effect of basic package price on subscriptions per 1,000 local residents using 2SLS.
 - b. Using your 2SLS results, what is a 99% confidence interval for the effect of the basic package price on subscriptions per 1,000 local

residents?

- c. Is there reason for concern about a weak instrument? Explain.

Dataset available at www.mhhe.com/prince1e

12. You are working as an analyst for a chain grocery store. The store uses “member cards” to keep track of customer purchases and sends coupons periodically to customers via e-mail. The chain is seeking to learn the impact of sending coupons for its generic-brand cereal on customer demand for that product. You have data for many customers spanning many weeks on number of generic-brand cereal boxes purchased and whether the customer received a coupon that week. Use the data in *Chap8Prob1213.xlsx* for this question. (LO5)

- a. Based on the data provided, write out an expression for the fixed effects model for the number of generic-brand cereal boxes purchased.
- b. Estimate your fixed effects model using dummy variable estimation.
- c. Interpret the coefficient estimate for the variable indicating whether a coupon was received.
- d. Why might excluding the fixed effects in your regression lead to a biased estimate for the effect of the coupon?

Dataset available at www.mhhe.com/prince1e

13. Refer to Problem 12. Again use the data in *Chap8Prob1213.xlsx* for this question. (LO6)
- a. Write out a reformulation of your fixed effects model using variables allowing for within estimation.
 - b. Estimate your reformulated fixed effects model using within estimation.
 - c. Build a 99% confidence interval for the effect of sending coupons on sales of generic-brand cereal boxes.

Prediction for a Dichotomous Variable

9

LEARNING OBJECTIVES

After completing this chapter, you will be able to:

- LO9.1** Identify a limited dependent variable and its applications.
- LO9.2** Describe the linear probability model.
- LO9.3** Identify merits and shortcomings of the linear probability model in practice.
- LO9.4** Model probit and logit models as determined by the realization of a latent variable.
- LO9.5** Calculate marginal effects for logit and probit models.
- LO9.6** Execute estimation of a probit and logit model via maximum likelihood.
- LO9.7** Identify the merits and shortcomings of probit and logit models in practice.

dataCHALLENGE *Changing the Offer to Change Your Odds*

Last year, a large accounting firm in Chicago, Northwest Accounting, made 38 entry-level accountant offers. [Table 9.1](#) below contains information on the salaries offered and whether the applicant accepted the offer.

TABLE 9.1 Employment Offers and Acceptances for Northwest Accounting

OFFER AMOUNT	ACCEPT?	OFFER AMOUNT	ACCEPT?
\$66,780	Accept	\$91,925	Accept
66,722	Accept	84,810	Accept
85,633	Decline	69,647	Accept
94,684	Decline	78,864	Decline
72,207	Accept	56,625	Decline
59,519	Decline	80,904	Decline
53,468	Accept	56,372	Accept
54,309	Decline	77,469	Accept
76,772	Decline	69,452	Accept
54,443	Decline	69,692	Decline
93,211	Accept	54,569	Decline
84,494	Accept	77,722	Decline
66,468	Decline	68,964	Accept
77,855	Accept	85,541	Accept
80,815	Decline	72,424	Decline
85,011	Accept	64,298	Accept
50,112	Decline	68,712	Accept
91,487	Decline	65,817	Decline
81,492	Accept	53,725	Decline

259

Northwest has hired you to learn if and how changes in the salaries it offers affect the likelihood of an applicant accepting the offered position. For example, the company would like to know how much more likely it would be that an applicant will accept the position when the offer is increased by \$5,000.

How could these data be used to estimate such a relationship?

Introduction

To this point in the book, every model we've studied has implicitly assumed that the dependent variable (or outcome) can take on essentially any value. Or at the very least, we have assumed that, even if our dependent variable does have limitations, they are not binding in practice. For example, when we use Sales as our dependent variable, we know it cannot be less than zero. However, our data never contained Sales figures at or very near zero, so this limitation was not consequential to our analysis.

In this chapter, we consider situations in which the dependent variable has limitations that are consequential. We discuss limited dependent variables in general, and then focus our attention on the most extreme form of limitation for a dependent variable (a dichotomy). We then detail alternative, highly utilized models for dependent variables that can take on only two values. The first is the *linear probability model*, which directly follows the regression models we've considered thus far in the book. The second and third models—the probit and logit models—are closely related. Both take a different approach toward modeling and estimation that specifically caters to dependent variables that are dichotomous. In describing these models, we highlight their merits and shortcomings, as pertains to making meaningful predictions about the effect of treatments on outcomes.

Limited Dependent Variables

260

LO 9.1 Identify a limited dependent variable and its applications.

A **limited dependent variable** is a dependent variable whose range of possible values has consequential constraints. Put another way, a limited dependent variable is unable to take on at least some values, and this limitation has “consequence.”

limited dependent variable A dependent variable whose range of possible values has consequential constraints.

Fully detailing exactly when a constraint on a dependent variable has consequence and when it does not can be a daunting, and tedious, task. Instead, we can provide the intuition as to when this is the case via two contrasting examples—one in which the dependent variable's constraints are not consequential and the other in which they are. First, consider a random variable Profits_i , representing the profits for firm i in a given year. Given no other information, Profits_i could take on a wide range of (both positive and negative) values. However, it is reasonable to believe there are bounds on these values: It is unimaginable for a firm's profits to exceed the entire GDP of the United States, as a gain or loss. Consequently, it is reasonable to impose a limitation on the variable Profits_i that it cannot exceed, say, \$20 trillion in absolute value. That would mean we have \$20 trillion as an upper bound and -\$20 trillion as a lower bound on Profits_i . These bounds are constraints on our Profits_i variable, but here we would argue these constraints have no apparent, meaningful consequence. These bounds could be consequential only if we observed Profits that approached them or if we tried to make predictions about Profits that came near them. Neither occurrence is at all realistic in practice.

In contrast to our Profits example, consider a random variable Spend_{it} representing the amount of money spent buying products online by household i in week t . Products essentially always have non-negative prices, so this random variable is constrained to be at least zero. Further, this constraint could be consequential. Consider the hypothetical data on Spend_{it} contained in [Table 9.2](#). Here, we see several observations with Spend_{it} values exactly equal to the constraint of zero. This suggests that the constraint is meaningful

in some way, because we see many instances of households choosing its exact value. A common way of incorporating the meaning of this constraint when modeling Spend_{it} as a dependent variable is to treat the realization of Spend_{it} as the result of two decisions: (1) The decision whether to buy *anything* online this week, and (2) if the household chooses to buy online this week, the decision of how much to spend. By building a model that incorporates this constraint at zero, we can better understand and predict the clustering of zeros we observe (and are likely to continue to observe) in the data.

There are many types of limited dependent variables, with various types of constraints. Some standard constraints include upper and/or lower bounds—for example, our Spend_{it} variable had a lower bound. Some constraints include the ability to take on only (typically highly limited) discrete values. As an example of that type of constraint, consider the variable Transit_i , which indicates person i 's primary method of commuting to and from work. Define Transit_i to be a discrete random variable limited to just five values: 1 if by car, 2 if by train, 3 if by bus, 4 if by walking, and 5 if other. If we attempted to model Transit_i as our dependent variable, it would certainly fall in the class of limited dependent variables since its constraints are clearly “consequential.”

Thus far in this section, we have provided a high-level overview of limited dependent variables. Ideally, we would detail each different type of limited dependent variable and associated models that can account for their limitations. Such a task would be quite long and often tedious; it could even warrant its own separate book. While we cannot detail

TABLE 9.2 Household Weekly Online Expenditures

HOUSEHOLD	WEEK	SPEND
1	1	\$214.87
1	2	0
1	3	103.95

1	4	0
2	1	0
2	2	47.88
2	3	32.19
2	4	85.32
3	1	152.34
3	2	0
3	3	105.58
3	4	0
4	1	243.11
4	2	65.48
4	3	0
4	4	172.45
5	1	56.78
5	2	0
5	3	0
5	4	172.67

all limited dependent variables, we focus the remainder of this chapter on one in particular that is the most limited and also likely the most utilized—a dichotomous (or binary) dependent variable.

A **dichotomous (or binary) dependent variable** is a limited dependent variable that can take on just two values, typically recorded as 0 and 1. You may also sometimes see the term “dummy dependent variable” used as synonymous with a dichotomous dependent variable. We refrain from using these terms interchangeably here, as dummy variables are often referenced in the context of independent (rather than dependent) variables. As a prime example, dummy variable estimation, detailed in [Chapter 8](#), refers to dummy variables being used as independent variables in the form of fixed effects.

dichotomous (or binary) dependent variable A limited dependent variable that can take on just two values, typically recorded as 0 and 1.

Dichotomous dependent variables measure many different types of outcomes. Among many other outcomes, they can measure, for example: purchase/don’t purchase, project success/project failure,

employed/unemployed, bankrupt/not bankrupt, approve/disapprove. To add further context, consider the following simple example. Suppose an upstart firm, called SaferContent, has developed new proprietary software designed to effectively protect clients' digital content on all of their devices (computer, smartphone, etc.). The software is available only through the firm's website, and purchase is in the form of a monthly subscription. An outcome variable of particular interest to SaferContent is the

262

purchase decision of individuals who visit the website. Specifically, they would like to gain a better understanding of the factors that influence whether a visitor to the website ultimately makes a purchase. The goal is to make sound predictions about the effects of changing strategic variables (e.g., price) on the purchase outcome.

Because the purchase decision is dichotomous (purchase/don't purchase) rather than quantitative (how much to purchase), the analysis conducted to gain this understanding will require the use of a dichotomous dependent variable. Specifically, SaferContent may collect data over the course of one day on visitors to its website along with the website's features (including price charged). In doing so, it can define the dichotomous dependent variable, $Purchase_i$, as:

$$Purchase_i = \begin{cases} 1 & \text{if visitor } i \text{ makes a purchase} \\ 0 & \text{if visitor } i \text{ does not make a purchase} \end{cases}$$

In the context of this dichotomous dependent variable, SaferContent would like to learn about factors that cause it to change from a 0 to a 1. For example, it may want to ask what

COMMUNICATING DATA 9.1

HOW TO MODEL AND PREDICT CORD CUTTING

A phenomenon of keen interest to those working in the television industry is that of “cord cutting,” in which subscribers to multichannel video programming distributors (“MVPDs,” such as distributors of cable and satellite television) discontinue their subscriptions. Industry analysts are interested in modeling and predicting the decision of households to cord cut. This process begins by defining the dependent variable and recognizing its limitations. Here, the dependent variable is the act of cord cutting, which we can define as a dichotomous dependent variable as follows:

$$\text{CordCut}_i = \begin{cases} 1 & \text{if household } i \text{ discontinues its subscription television} \\ 0 & \text{if household } i \text{ does not discontinue its subscription television} \end{cases}$$

Analysts may collect data on CordCut for a sample of households who had subscription television service in, say, 2017, by then observing whether those households had subscription television service in 2018 as well.

Having established the dependent variable/outcome they'd like to predict, analysts may then consider factors that might influence the realized outcome for CordCut. One such factor might be availability of on-demand television provided by the MVPDs. If we had a measure of on-demand availability, call it OnDemand, and it varied perhaps across time and regions, we could try to measure its impact on cord cutting. However, similar to our SaferContent example in the text, doing so raises several questions, such as:

- How do we model the relationship between CordCut and OnDemand?
- How do we interpret predictions that involve a fractional change in CordCut, or do we somehow force those not to occur?
- Can we be sure predicted effects of OnDemand will not lead to unrealistic predictions for values of CordCut, such as those falling outside of 0 and 1?

We address these and other questions throughout the remainder of the

is the response of Purchase to a \$10 decrease in the monthly subscription fee for its software. Even from just posing this simple question, it becomes apparent that the limitation of Purchase to being just 0 or 1 is consequential: Since Purchase can only take on two values, is the effect of a price change limited to being just 0 or 1? If not, how do we interpret a fractional response (e.g., 0.3)? And, is there an interpretation for a response that is bigger than one in absolute value (e.g., 1.4 or -1.7)?

Answering questions about factors affecting Purchase will require us to assume a data-generating process, as we've done for similar problems throughout the book. No matter the model we choose for the data-generating process of a dichotomous dependent variable, recognizing the variable's limitations is always important for interpretation of the corresponding estimates. Also, it is often important to specifically tailor our model to the limitations of the dependent variable in order to get sensible predictions, particularly around the constraint(s).

In the next section, we discuss the case in which we do not tailor our model of the data-generating process to account for the dichotomous nature of the dependent variable—that is, we proceed with a standard regression model with no constraints, as in the prior chapters. We then highlight how to properly interpret the findings, the merits of taking this “standard” approach, and situations in which this approach risks nonsensical predictions (among other issues). Then, in the final section, we discuss models of the data-generating process that specifically account for dichotomy in the dependent variable. For these models, we again highlight how to interpret their findings properly, the merits of taking these alternative approaches, and some of their shortcomings.

The Linear Probability Model

In this section, we define and interpret the *linear probability model*—a widely utilized model for dichotomous dependent variables. We detail the merits of this model in practice, but then conclude by discussing its shortcomings and when they are particularly likely to be consequential. The next section introduces alternative models for a dichotomous dependent variable, which can overcome some of the shortcomings of the linear probability model.

DEFINITION AND INTERPRETATION

LO 9.2 Describe the linear probability model.

Consider again our SaferContent example: Suppose we are interested in measuring the effect of the subscription fee on the decision to make a purchase. The purchase decision is our dependent variable, and we know it is dichotomous. However, suppose we ignore the fact that Purchase is dichotomous and treat it like all other dependent variables we've analyzed in prior chapters. In that case, we would simply assume a data-generating process for Purchase, and use regression analysis to estimate the parameters for that process. Specifically, we might assume the following data-generating process for the purchase of a SaferContent subscription:

$$\text{Purchase}_i = \alpha + \beta \text{SubFee}_i + U_i$$

Here, SubFee_i is the subscription fee faced by individual i .

Suppose we attempt to estimate the parameters of our assumed data-generating process using regression analysis. We would be utilizing a **linear probability model**, defined

linear probability model Regression analysis applied to a dichotomous dependent variable.

as regression analysis applied to a dichotomous dependent variable. Our definition of a linear probability model, *per se*, does not require an interpretation of causality. The act of fitting the equation $\text{Purchase} = \alpha + \beta\text{SubFee}$ to the data by solving the moment conditions is an application of a linear probability model, and as we know from [Chapter 6](#), this process alone does not imply causality. However, when we make the necessary additional assumptions for causality (the function we fit to the data is a determining function in a data-generating process, etc.), we can make causal inferences as we did for regression analysis with unrestrained dependent variables.

To better understand the linear probability model, consider an associated dataset and the estimates from a simple regression. In [Table 9.3](#), we present data for our SaferContent example, and in [Table 9.4](#), we present regression estimates that fit the function $\text{Purchase} = \alpha + \beta\text{SubFee}$ to the data.

TABLE 9.3 Data on Subscription Fees and Purchase Decisions for SaferContent

INDIVIDUAL	PURCHASE	SUBFEE	INDIVIDUAL	PURCHASE	SUBFEE	IND
1	0	19.08	26	0	25.87	
2	0	21.19	27	0	16.6	
3	1	20.87	28	0	27.41	
4	0	27.62	29	0	29.28	
5	1	26.95	30	0	16.33	
6	0	26.44	31	0	15.94	
7	1	16.65	32	1	19.66	
8	0	23.89	33	0	26.26	
9	1	19.29	34	1	28.92	
10	0	22.84	35	1	21.11	
11	0	28.87	36	1	17.62	
12	1	22.39	37	0	20.27	
13	1	15.6	38	0	20.89	
14	0	23.85	39	1	20.51	
15	0	20.16	40	1	17.14	
16	0	21.89	41	1	17.54	
17	1	21.68	42	0	20.22	

18	0	25.14	43	0	18.16	
19	1	24.44	44	0	20.13	
20	1	16.36	45	1	18.82	
21	1	18.09	46	1	21.07	
22	1	16.77	47	1	25.08	
23	1	18.87	48	0	23.55	
24	0	22.37	49	0	29.87	
25	0	29.29	50	1	21.35	

265

TABLE 9.4 Regression Results for Purchase Regressed on Subscription Fee

	COEFFICIENTS	STANDARD ERROR	t STAT	P-VALUE	LOW
Intercept	1.65429339	0.294751275	5.612506305	3.38971E-07	1.066
SubFee	-0.052585324	0.012985261	-4.049616232	0.000126414	-0.07

Based on the estimates in [Table 9.4](#), our best guess for the determining function for Purchase would be: $\text{Purchase}_i = 1.65 - 0.05 \times \text{SubFee}_i$. If Purchase were not a limited dependent variable, interpreting this determining function would be straightforward. If we made the assumptions that establish causality (see [Reasoning Box 6.5](#)), we would simply interpret this result as indicating that a \$1 increase in subscription fee reduces Purchase by 0.05, on average.

Because Purchase is a dichotomous dependent variable, our interpretation of the effect of SubFee differs from the “standard” (unlimited dependent variable) case. To see how, first note that Purchase can take on only the values of 0 and 1, so no single realization can change by this average effect. We can never see an individual's realization for Purchase change by 0.05—it could increase or decrease only by one. Contrast this with examples where the dependent variable is not limited—e.g., where we regress stores' monthly Profits on Price. In such a regression, we might find that Profits decline by \$5,238.41 with a \$1 increase in Price, on average. For this contrasting Profits example, it makes sense to consider a single store's monthly profits changing

by the average effect of a \$1 increase in Price, whereas this is not possible for our SaferContent example.

Given this restriction on our interpretation of the effect of SubFee, how do we interpret this effect for an individual? The answer to this question stems from how this effect is calculated in practice. Consider an alternative SaferContent dataset with just two values for SubFee (\$20 and \$19), as in [Table 9.5](#). We have four purchases (out of 20, i.e., 20%) at the price of \$19 and three purchases (out of 20, i.e., 15%) at the price of \$20. If we regress Purchase on SubFee, we again get a coefficient of -0.05 on SubFee. However, for this simplified dataset, it is clear exactly what this value represents: It captures that when price increased by \$1, the fraction of individuals purchasing declined by 5% (from 20% to 15%). When considering a single individual, we can interpret this as the probability of a purchase declining by 5% when the price increases by \$1. Hence, we can view 0.05 as an average change in the *probability* of a purchase with a \$1 price change.

We can extend our interpretation for the SaferContent example to general models with dichotomous dependent variables. Suppose we've assumed a data-generating process of

$$Y_i = \alpha + \beta_1 X_{1i} + \cdots + \beta_K X_{Ki} + U_i$$

and Y is a dichotomous dependent variable. We can interpret the β s as changes in the probability of Y taking on a value of 1. We express this a bit more formally, e.g., for β_1 , as:

$$\beta_1 = \Pr(Y = 1|X_1 + 1, X_2, \dots, X_K) - \Pr(Y = 1|X_1, \dots, X_K)$$

In words, each β represents the change in probability of Y equaling 1 when its corresponding X increases by 1, holding all the other X s constant. It is this

interpretation that motivates the name we apply to such analysis: the linear *probability* model. We summarize these general points in [Reasoning Box 9.1](#).
266

TABLE 9.5 SaferContent Data with Subscription Fees of only \$19 and \$20

PURCHASE	SUBFEE	PURCHASE	SUBFEE
0	20	0	19
0	20	1	19
0	20	0	19
1	20	0	19
0	20	0	19
0	20	1	19
0	20	0	19
0	20	0	19
0	20	0	19
1	20	0	19
0	20	1	19
0	20	0	19
0	20	0	19
0	20	0	19
0	20	0	19
0	20	1	19
1	20	0	19
0	20	0	19

REASONING BOX 9.1

**INTERPRETATION OF A
LINEAR PROBABILITY
MODEL**

IF:

1. We assume the data-generating process for Y to be:

$$Y_i = \alpha + \beta_1 X_{1i} + \cdots + \beta_K X_{Ki} + U_i$$

2. Y is a dichotomous dependent variable

THEN:

β_1, \dots, β_K represent the change in the probability of Y equaling one with a one unit increase in X_1, \dots, X_K (respectively), holding all other X s constant.

For example, we can express β_1 as:

$$\beta_1 = \Pr(Y = 1|X_1 + 1, X_2, \dots, X_K) - \Pr(Y = 1|X_1, X_2, \dots, X_K).$$

MERITS AND SHORTCOMINGS

LO 9.3 Identify merits and shortcomings of the linear probability model in practice.

The linear probability model represents one approach we may employ when attempting to make predictions for dichotomous dependent variables. In the next section, we consider some alternative approaches, but before doing so, it is useful to highlight the merits and shortcomings of the linear probability model. The latter will help motivate our subsequent consideration of alternatives.

The merits of the linear probability model are easy to summarize. The linear probability model imposes no restrictions on the associated regression analysis, so all methods discussed in [Chapters 7 and 8](#) (use of dummy variables, selecting controls, instrumental variables, panel data methods) seamlessly apply. As we've noted, the interpretation of the estimates will differ (they represent changes in probabilities), but all else is identical.

The merits of the linear probability model stem from the fact that, other than through interpretation, it ignores the limitation of the dependent variable.

However, ignoring that the dependent variable can take on only values of 0 and 1 can be consequential, and in such cases, represents important shortcomings of the linear probability model.

The first shortcoming has to do with interpretation of the data-generating process as a whole. Consider again our SaferContent example and the associated data-generating process:

$$\text{Purchase}_i = \alpha + \beta \text{SubFee}_i + U_i$$

Suppose we knew that $\alpha = 0.8$ and $\beta = -0.02$, so the data-generating process is

$$\text{Purchase}_i = 0.8 - 0.02 \text{SubFee}_i + U_i$$

Notice that the fact Purchase can be only 0 or 1 has very stark implications about the distribution of the unobservables (U). To see this, consider the possible values of U when the subscription fee is \$20. Here, we have $\text{Purchase} = 0.8 - 0.02(20) + U = 0.4 + U$. Therefore, since Purchase can be only 0 or 1, U can be only 0.6 or -0.4. Broadly speaking, for any given subscription fee, there are only two values for the unobservables that produce a feasible value for Purchase.

This severe limitation on the inferred distribution of U has two key implications. The first is technical: It is not possible for the unobservables to be homoscedastic (have constant variance). Recall from [Chapter 6](#) that homoscedasticity was one of the assumptions allowing us to build confidence intervals and conduct hypothesis tests using our regression results. Proving the unobservables cannot be homoscedastic is outside the scope of this book; however, we note here that one can make relatively easy corrections to the standard errors when this assumption is violated. Therefore, this implication has minimal consequence and so is not a very serious shortcoming of the

linear probability model in practice.

The second implication centers on our conceptualization of the unobservables. Recall that the unobservables (U) represent all other factors that affect the outcome. For our SaferContent example, these factors may comprise of the individual's age, education, income, etc. Conceptually, the effects of all these factors combine to generate a realization of U ; however, it is difficult to envision how, for a given subscription fee, the combination of these factors would always result in only one of two net effects. Consider again the scenario where $\text{Purchase}_i = 0.8 - 0.02\text{SubFee}_i + U_i$, and now we have a group of individuals facing a subscription fee of \$20. All of these individuals may vary extensively in their ages,

268

education, income, etc., but according to our data-generating process, the combined effects of these variables is always equal to 0.6 or -0.4. This limitation on the combined effects of unobserved factors does not compromise our ability to get consistent estimates for the parameters of the determining function. However, it does represent a shortcoming of the linear probability model whenever we try to make theoretical claims about the unobservables and their combined effects. For example, suppose we believed income was the only unobserved factor affecting a Purchase ($U = f(\text{Income})$). Then, according to our model, the effect of Income on Purchase is always 0.6 or -0.4 when the subscription fee is \$20, always 0.4 or -0.6 when the subscription fee is \$30, etc. While technically possible, it is difficult to conjure a sound rationale as to why the unobservables would work in this way.

The second shortcoming of the linear probability model that we highlight has to do with the lack of restrictions on the range of predicted values for the outcome. In our SaferContent example, using the data in [Table 9.2](#), we estimated the determining function to be: $\text{Purchase}_i = 1.65 - 0.05 \times \text{SubFee}_i$. Now, suppose the current subscription fee is \$25 and the observed subscription rate at that price is 32%. We might ask what would happen if we tried increasing the price by \$10 to \$35. According to our estimates, we

would predict that a \$10 price increase would lower the probability of a purchase by $0.05 \times 10 = 0.50$, or 50%. This means we'd predict the purchase rate would drop from 32% to -18%. Of course, it is not possible for a probability to be outside of 0–100%; but because no restrictions were placed on our model, it is capable of making impossible predictions, such as a negative probability.

The above prediction for a \$10 increase in subscription fee illustrates how linear probability models are capable of making a **limit-violating prediction**, defined as a predicted value for a limited dependent variable that does not fall within that variable's limits. Hence, the second, and most glaring, shortcoming of the linear probability model is that it can lead to limit-violating predictions.

limit-violating prediction A predicted value for a limited dependent variable that does not fall within that variable's limits.

We conclude by elaborating on the extent to which limit-violating predictions are a problem for the linear probability model. Consider the general formulation of a linear probability model, where Y_i takes on the values of 0 or 1:

$$Y_i = \alpha + \beta_1 X_{1i} + \cdots + \beta_K X_{Ki} + U_i$$

First note that, for many applications, limit-violating predictions may not be a problem in practice. This is because there are often practically imposed limitations on the X s that preclude predictions for Y outside of the range 0–1. In our SaferContent example, the observed subscription fees all ranged between \$15 and \$30. Given our estimated determining function of $\text{Purchase}_i = 1.65 - 0.05 \times \text{SubFee}_i$, we generally will not have predictions for the probability of a purchase that fall outside of 0.15 ($1.65 - 0.05 \times 30$) and 0.90 ($1.65 - 0.05 \times 15$) if we limit our predictions to be for subscription fees in the

observed range. Of course, we will certainly encounter limit-violating predictions if we consider subscription fees outside of \$15–\$30, but we should be wary of making such predictions for reasons beyond just the possibility of violating limits of our model (as we detail in [Chapter 10](#)).

Our second key point with regard to limit-violating predictions concerns whether we could engineer the X s in such a way as to preclude predictions for Y outside of 0–1. Recall that for linear regression, which includes the linear probability model, the X s could be different functions of the same variable ($X_1 = \text{Price}$, $X_2 = \text{Price}^2$, etc.). In our SaferContent

269

example, we could try to include different functions of the subscription fee in an attempt to limit the range of values for Purchase the determining function might produce. We may start with a simple approach by assuming

$$\text{Purchase}_i = \alpha + \beta \left(\frac{1}{\text{SubFee}_i} \right) + U_i$$

This alternative functional form may seem helpful toward limiting the associated values for Purchase, since $\left(\frac{1}{\text{SubFee}} \right)$ will be between 0 and 1 for any subscription fee more than \$1. However, there is no restriction on β (or α), so we could still end up with limit-violating predictions, e.g., predictions above 1 for Purchase if β is a large number.

9.1

Demonstration Problem

An outcome variable, Y , can take on only the values 0 or 1. In attempting to measure the effects of X_1 and X_2 on Y , you've collected a sample of size 200 on these three variables, and assumed the following:

- A. The data-generating process for Y is: $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + U_i$
- B. $\{Y_i, X_{1i}, X_{2i}\}_{j=1}^{200}$ is a random sample
- C. $E[U] = E[U \times X_1] = E[U \times X_2] = 0$

You regress Y on X_1 and X_2 , which yields the results in [Table 9.6](#).

TABLE 9.6 Regression Results for Y Regressed on X_1 and X_2

	COEFFICIENTS	STANDARD ERROR	t STAT	P-VALUE	LOW
Intercept	-0.02100635	0.052861193	-0.397386983	0.691512341	-0.02100635
X1	0.027106899	0.003688685	7.34866151	5.2088E-12	0.01
X2	0.030375547	0.004368002	6.954106077	5.10106E-11	0.02

- Interpret the estimates for β_1 and β_2 .
- Why are the p -values and confidence intervals associated with β_1 and β_2 invalid?
- If we believe values for X_1 will always be between 10 and 20 and values for X_2 will always be between 3 and 8, is there reason for concern about limit-violating predictions?

Answer:

- The estimate for β_1 implies that, when X_1 increases by one unit and X_2 is held constant, the probability of Y equaling 1 increases by 2.7 percentage points. The estimate for β_2 implies that, when X_2 increases by one unit and X_1 is held constant, the probability of Y equaling 1 increases by 3.0 percentage points.
- These statistics are invalid because the assumption of homoscedasticity is violated for the linear probability model. We must make corrections to the standard errors of the estimators in order to conduct hypothesis tests and/or build confidence intervals.
- No. The range of values for the determining function is approximately 0.34 ($= -0.021 + 0.027 \times 10 + 0.030 \times 3$) and 0.76 ($= -0.021 + 0.027 \times 20 + 0.030 \times 8$), which is within the range of 0 to 1. Consequently, there should

not be concerned about limit-violating predictions for this model with these data.

270

There is no way to construct the determining function to ensure the dependent variable always will be between 0 and 1 using the linear probability model. However, if we are willing to consider alternative models—models that are not linear in the parameters—we can effectively impose this constraint. We introduce the two most common of these alternative models in the next section.

COMMUNICATING DATA 9.2

CHARACTERIZING ENDOGENEITY WITHIN A LINEAR PROBABILITY MODEL

Let's revisit our cord-cutting example from [Communicating Data 9.1](#). Suppose we assume the following data-generating process for the decision to cord the cut:

$$\text{CordCut}_i = \alpha + \beta \text{OnDemand}_i + U_i$$

Recall that CordCut_i equals 1 if household i discontinues its subscription television between 2017 and 2018, and 0 otherwise. Further, suppose that OnDemand_i is the number of hours of television that is available to household i from its subscription television provider entering 2018.

We know from our previous discussion that the unobservables (U) can take on only two values (the value that makes CordCut equal 1 and the value that makes CordCut equal 0) for any given value of OnDemand . This severe limitation on the possible values for U affects how we characterize an endogeneity problem, if one exists. In our CordCut example, an endogeneity

problem exists if the unobservables are correlated with OnDemand. If $\alpha + \beta$ OnDemand generally lies between 0 and 1, then each realization of U is either positive (which makes CordCut equal 1) or negative (which makes CordCut equal 0). Hence, an important component of endogeneity within the linear probability model is the frequency of “good” versus “bad” draws for the unobserved factors. For example, we might worry that our estimate for β will suffer from an endogeneity problem if factors leading to cord cutting (e.g., job loss) are particularly likely for individuals with high amounts of OnDemand hours available.

Probit And Logit Models

Motivated by the limitations of the linear probability model, this section presents two closely related alternative models for dichotomous dependent variables. Both are designed to overcome the limitations of the linear probability model. We explain in detail how both models are formulated and estimated, and we discuss their merits and shortcomings. We note that the choice between a linear probability model and the alternative models we present in this section is not an obvious one—there is no universally “right” model. Rather, it is important to be aware of the merits and shortcomings of each modeling option, and be considerate of them when obtaining and interpreting estimates for a treatment effect. Results for a model of a dichotomous dependent variable are generally most convincing when they are at least qualitatively consistent across the linear probability model and the popular alternatives we present as follows.

271

LATENT VARIABLE FORMULATION

LO 9.4 Model probit and logit models as determined by the realization of a latent variable.

The key difference between the linear probability model and the models we introduce in this subsection is the connection between the determining function, the unobservables, and the dependent variable. Consider our SaferContent example and its formulation when applying a linear probability model:

$$\text{Purchase}_i = \alpha + \beta \text{SubFee}_i + U_i$$

We have that the realization of a purchase (0 or 1) is literally the sum of the realized value of the determining function ($\alpha + \beta \text{SubFee}_i$) plus the realized value of the unobservables (U_i). As we noted at the end of the last section, this model suffers from two shortcomings: (1) It is hard to believe the determining function and unobservables always add up to exactly 0 or 1, and (2) predictions about the effect of the subscription fee on purchases may be unrealistic (e.g., imply a change in the probability of a purchase of more than 100%).

Let's now reconsider the relationship among the determining function, unobservables, and the dependent variable. Rather than set the dependent variable equal to the sum of the determining function and unobservables, suppose we let the value of the dependent variable depend on this sum, but only in a coarse way. We define the sum of the determining function and unobservables as equaling a latent variable, and then let the value of the dependent variable depend on whether the latent variable is positive or not. A **latent variable** is a variable that cannot be observed, but information about it can be inferred from other observed variables (e.g., the dependent variable).

latent variable A variable that cannot be observed, but information about it can be inferred from other observed variables.

We can illustrate the idea of latent variables in this context using our SaferContent example: We define the sum of the determining function ($\alpha +$

βSubFee_i) and unobservables (U_i) to be a latent variable. To make this approach intuitive and grounded in economic theory, let's call the latent variable Utility. Thus, we have $\text{Utility}_i = \alpha + \beta \text{SubFee}_i + U_i$. We assume a Purchase occurs ($\text{Purchase}_i = 1$) if Utility_i is positive (> 0) and a Purchase does not occur ($\text{Purchase}_i = 0$) if Utility_i is not positive (≤ 0). Hence, we can express the purchase decision as:

$$\text{Purchase}_i = \begin{cases} 1 & \text{if } \text{Utility}_i > 0 \\ 0 & \text{if } \text{Utility}_i \leq 0 \end{cases}$$

Or, equivalently:

$$\text{Purchase}_i = \begin{cases} 1 & \text{if } \alpha + \beta \text{SubFee}_i + U_i > 0 \\ 0 & \text{if } \alpha + \beta \text{SubFee}_i + U_i \leq 0 \end{cases}$$

In our SaferContent example, we do not observe Utility, but the realized value of Purchase tells us something about Utility, thus making it satisfy the definition of a latent variable. If $\text{Purchase} = 1$, we know Utility is greater than 0; if Purchase is 0, we know Utility is less than or equal to 0. Also, note that the determining function now directly determines values for Utility and only indirectly determines values for the dependent variable, Purchase.

Given this latent variable formulation of the data-generating process for Purchase_i , it can be helpful toward building intuition if we summarize it all in words: Here, we

TABLE 9.7 Examples of Dichotomous Dependent Variables Coupled with Latent Variables

DEPENDENT VARIABLE	LATENT VARIABLE

Make a Purchase	Utility
Open a New Business	Profits
Hire a New Employee	Net Revenues
Terminate Subscription	Utility
Acquire a Competing Firm	Profits

have defined the Utility individual i receives from purchasing SaferContent software as depending on the subscription fee and other factors. If the utility of purchasing is strictly positive, that individual makes a purchase; if the utility of purchasing is negative or zero, that individual does not make a purchase. Essentially, we have added another layer to the purchasing decision: The determining function and unobservables determine Utility, and then Utility (positive or negative) determines the purchase decision.

Modeling a dichotomous outcome as the result of a latent variable crossing (or not crossing) a threshold (e.g., zero) has many intuitive applications. In our SaferContent example, we have the outcome of a purchase depending on whether utility is “high enough.” Some other applications are summarized in [Table 9.7](#) (by no means an exhaustive list).

Now consider a generalization of this way of modeling a dichotomous dependent variable. Define our latent variable, Y_i^* , as the sum of the determining function and the unobservables:

$$Y_i^* = \alpha + \beta_1 X_{1i} + \cdots + \beta_K X_{Ki} + U_i$$

Then, we define the realization of our dependent variable, Y_i , to be 1 if the latent variable exceeds 0, and 0 otherwise:

$$Y_i = \begin{cases} 1 & \text{if } Y_i^* > 0 \\ 0 & \text{if } Y_i^* \leq 0 \end{cases}$$

Notice that the latent variable formulation for Y overcomes our first criticism for the linear probability model. Expressed this way, we do not need the determining function and unobservables to always add up exactly to 0 or 1. We've placed no restriction on this sum, and by simply comparing it to 0, we always get a value for Y of 0 or 1.

The latent variable formulation for Y also prevents unreasonable predictions about the probability of Y equaling 1 (our second criticism of the linear probability model). To see this, first note that we can easily express the probability of Y equaling 1 for any given values for the X s:

$$\Pr(Y_i = 1 | X_{1i}, \dots, X_{Ki}) = \Pr(Y^* > 0 | X_{1i}, \dots, X_{Ki})$$

This equation states that the probability the outcome (Y) equals 1, given the values for the X s, is equal to the probability that the latent variable (Y^*) is greater than 0, given the values for the X s. Next, if we plug in our formula for the latent variable, we have:

$$\Pr(Y_i = 1 | X_{1i}, \dots, X_{Ki}) = \Pr(\alpha + \beta_1 X_{1i} + \dots + \beta_K X_{Ki} + U_i > 0 | X_1, \dots, X_K)$$

273

Lastly, note that with the values of the X s given, uncertainty about Y is completely due to uncertainty about U . Hence, it makes sense to isolate U within our probability function:

$$\Pr(Y_i = 1 | X_{1i}, \dots, X_{Ki}) = \Pr(U_i > -\alpha - \beta_1 X_{1i} - \dots - \beta_K X_{Ki} | X_1, \dots, X_K)$$

We can see how this latent variable formulation for Y overcomes the second criticism of the linear probability model. While the determining

function is unconstrained, the probability that Y equals 1 is explicitly defined to be a probability in terms of U , and so constrained to be between 0 and 1. Hence, even if the determining function becomes very large for some values of the X s, the probability of Y equaling 1 will just approach 1, and vice versa.

Since we have repurposed the determining function as directly determining a latent variable rather than the outcome itself, the parameters of the determining function (the β s) represent the change in the latent variable (Y^*)—not the outcome—when their corresponding X s increase by one, holding all the other X s constant. In our SaferContent example, β is the change in Utility when the subscription fee increases by \$1, and does not by itself tell us how the likelihood of a Purchase changes with the subscription fee. Often it is not our primary interest to assess the effect of an independent variable on the latent variable. We may have limited interest in how subscription fees affect Utility. Rather, we'd like to know the effect of the independent variables on the likelihood of the outcome.

Before assessing how independent variables affect the likelihood of an outcome using our latent variable formulation, we first must make an assumption about the distribution of U . There are two highly popular assumed distributions for U : The first is familiar—the standard normal distribution. We assume U_i is distributed as a normal random variable with mean 0 and standard deviation of 1, written as $U_i \sim N(0,1)$. A latent variable formulation for a dichotomous dependent variable that assumes a standard normal distribution for the unobservables is defined as a **probit model**.

probit model A latent variable formulation for a dichotomous dependent variable that assumes a standard normal distribution for the unobservables.

Probabilities for the probit model simply utilize our knowledge of the normal distribution. In [Figure 9.1](#), we illustrate the probability that Y equals 1 for given values of the X s using the formula:

$$\Pr(Y_i = 1 | X_{1i}, \dots, X_{Ki}) = \Pr(U_i > -\alpha - \beta_1 X_{1i} - \dots - \beta_K X_{Ki})$$

Note in the graph that $\varphi(U)$ is the probability density function (pdf), for the standard normal distribution (discussed in [Chapter 3](#)).

274

FIGURE 9.1 Probability Y Equals 1 for Given Xs, Assuming Standard Normal Distribution for U

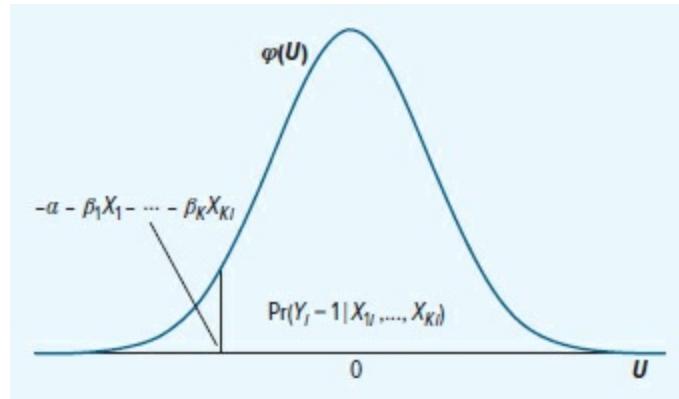
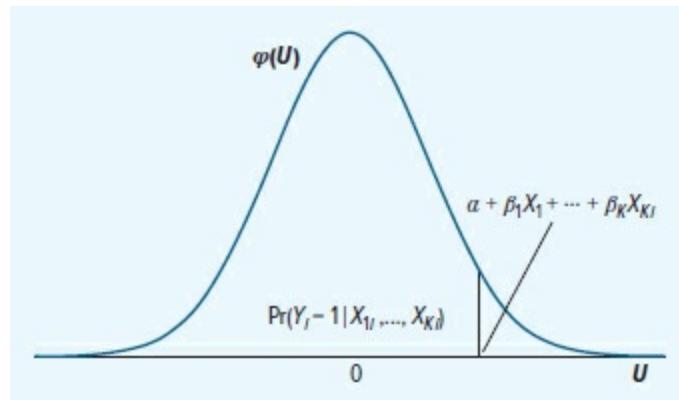


FIGURE 9.2 Probability Y Equals 1 for Given Xs, Assuming Standard Normal Distribution for U and Using cdf for U



Since we know U_i is a standard normal random variable, we can simplify the previous expression for $\Pr(Y_i = 1 | X_{1i}, \dots, X_{Ki})$ using this knowledge. First, we know the normal distribution is symmetric around its mean (which is 0 for U). Therefore, we know that:

$$\Pr(U_i > -\alpha - \beta_1 X_{1i} - \cdots - \beta_K X_{Ki} | X_{1i}, \dots, X_{Ki}) = \Pr(U_i < \alpha + \beta_1$$

Next, define $\Phi(\cdot)$ as the cumulative distribution function (cdf) for a standard normal random variable U , where $\Phi(m) = \Pr(U < m)$. Then, for the probit model, we have:

$$\Pr(Y_i = 1 | X_{1i}, \dots, X_{Ki}) = \Phi(\alpha + \beta_1 X_{1i} + \cdots + \beta_K X_{Ki})$$

We illustrate this idea in [Figure 9.2](#). Notice that this generates an identical probability (same area under the standard normal distribution) as in [Figure 9.1](#).

The second popular assumption for the distribution of U is the logistic distribution. Just like the normal, there are many variations of the logistic distribution, but the typical assumption is that $U_i \sim \text{Logistic}(0,1)$. This distribution has a pdf of:

$$f(z) = \frac{e^{-z}}{(1 + e^{-z})^2}$$

A latent variable formulation for a dichotomous dependent variable that assumes a Logistic(0,1) distribution for the unobservables is defined as a **logit model**.

logit model A latent variable formulation for a dichotomous dependent variable that assumes a Logistic(0,1) distribution for the unobservables.

The logistic distribution is not nearly as well known as the normal distribution, so why is it a popular choice for the distribution of the

unobservables? The answer is that the logistic distribution generates a simple formula for the probability of Y equaling 1 for a given set of X s. Contrast this feature with the probit model, where we rely on a computer to generate various probabilities for the normal distribution. When we assume that $U_i \sim \text{Logistic}(0,1)$, the probability that Y equals one for given values of the X s can be expressed as:

$$\Pr(Y_i = 1 | X_{1i}, \dots, X_{Ki}) = \frac{1}{1 + e^{-\alpha - \beta_1 X_{1i} - \dots - \beta_K X_{Ki}}}$$

The logistic distribution looks very similar to the normal (bell-shaped, covering the whole real line), so it is not a huge departure from the normal assumption and will result in similar probability calculations. However, its simplified formula for probabilities can facilitate exposition

275

and more complex analyses beyond simple probability calculations. For the remainder of the chapter, we will present results using both models, and the similarities will be apparent.

Why do analysts choose one model over the other? It is difficult to give an exhaustive answer. However, a simplified answer is the probit model assumes a familiar distribution that is known to often occur “naturally” in many settings, whereas the logit model assumes a distribution that resembles the normal but has more desirable formulaic features (e.g., probability formulas). We summarize the key differences between the probit and logit model in [Reasoning Box 9.2](#).

REASONING BOX 9.2

CONTRASTING THE PROBIT AND LOGIT MODEL

Define a latent variable, Y_i^* , as the sum of a determining function and

unobservables:

$$Y_i^* = \alpha + \beta_1 X_{1i} + \cdots + \beta_K X_{Ki} + U_i$$

Define a dichotomous dependent variable, Y_i , to be 1 if the latent variable exceeds 0, and 0 otherwise:

$$Y_i = \begin{cases} 1 & \text{if } Y_i^* > 0 \\ 0 & \text{if } Y_i^* \leq 0 \end{cases}$$

IF:

The unobservables, U_i , are distributed standard normal, i.e., $U_i \sim N(0,1)$

THEN:

1. The distribution of Y_i satisfies the assumptions of the probit model.
 2. The probability that Y_i equals 1, given the values for X_1, \dots, X_K , can be expressed as:
-

$$\Pr(Y_i = 1 | X_{1i}, \dots, X_{Ki}) = \Phi(\alpha + \beta_1 X_{1i} + \cdots + \beta_K X_{Ki})$$

where $\Phi(\cdot)$ is the cumulative distribution function for a standard normal random variable. Alternatively,

IF:

The unobservables, U_i , are distributed Logistic(0,1)

THEN:

1. The distribution of Y_i satisfies the assumptions of the logit model.
 2. The probability that Y_i equals 1, given the values for X_1, \dots, X_K , can be expressed as:
-

$$\Pr(Y_i = 1 | X_{1i}, \dots, X_{Ki}) = \frac{1}{1 + e^{-\alpha - \beta_1 X_{1i} - \dots - \beta_K X_{Ki}}}$$

276

MARGINAL EFFECTS

LO 9.5 Calculate marginal effects for logit and probit models.

We have established how to calculate the probability of the outcome equaling 1 for both the probit and logit models. Such calculations were a means to an end, as our ultimate goal is to establish how *changes* in the independent variables affect the probability of the outcome equaling 1. For the linear probability model, the parameters of the determining function (the β s) directly measure this relationship—each represents the change in probability of the outcome equaling 1 when its corresponding X s increase by 1, holding all the other X s constant. Unfortunately, when we utilize a latent variable formulation, the parameters of the determining function do not have such a simple interpretation.

Before attempting to analyze how probabilities change with independent variables in probit and logit models, it is useful to establish a more general definition of this concept. Roughly speaking, we can define the rate of change in the probability of a dichotomous dependent variable equaling 1 with a one-unit increase in an independent variable (holding all other independent variables constant) as a **marginal effect**. Notice that for the linear probability model, the β s in the determining function measure marginal effects (see [Reasoning Box 9.1](#)). However, for the probit and logit models, the β s no longer have this interpretation.

marginal effect The rate of change in the probability of a dichotomous dependent variable equaling 1 with a one-unit

increase in an independent variable (holding all other independent variables constant).

To see how we calculate marginal effects for probit and logit models, consider again a general latent variable model. Let

$$Y_i^* = \alpha + \beta_1 X_{1i} + \cdots + \beta_K X_{Ki} + U_i$$

be the latent variable, and Y_i be 1 if the latent variable exceeds 0, and 0 otherwise. Then, according to our definition, the marginal effect of X_j is

$$\text{MargEff}_{xj} = \Pr(Y_i = 1 | X_{1i}, \dots, X_{ji} + 1, \dots, X_K) - \Pr(Y_i = 1 | X_{1i}, \dots, X_{ji}, \dots, X_K)$$

If we apply this definition specifically to the probit and logit models, we have the following.

For probit:

$$\begin{aligned} \text{MargEff}_{xj} &= \Phi(\alpha + \beta_1 X_{1i} + \cdots + \beta_j(X_{ji} + 1) + \cdots + \beta_K X_{Ki}) \\ &\quad - \Phi(\alpha + \beta_1 X_{1i} + \cdots + \beta_j X_{ji} + \cdots + \beta_K X_{Ki}) \end{aligned}$$

For logit:

$$\begin{aligned} \text{MargEff}_{xj} &= \frac{1}{1 + e^{-\alpha - \beta_1 X_{1i} - \cdots - \beta_j(X_{ji} + 1) - \cdots - \beta_K X_{Ki}}} \\ &\quad - \frac{1}{1 + e^{-\alpha - \beta_1 X_{1i} - \cdots - \beta_j X_{ji} - \cdots - \beta_K X_{Ki}}} \end{aligned}$$

Notice that in neither case does the marginal effect simplify to β_j , meaning we generally get different marginal effects for probit and logit models compared to the linear probability model.

We can easily illustrate the calculation of marginal effects using our SaferContent example. Suppose the subscription fee is \$22, and we increase the subscription fee by \$1. The change in the probability of a Purchase (the marginal effect) is:

$$\Pr(\text{Purchase}_i = 1 | \text{SubFee}_i = \$23) - \Pr(\text{Purchase}_i = 1 | \text{SubFee}_i = \$22)$$

To calculate the marginal effect, we must know the values for α and β , and we must assume a logit or probit model. Suppose we knew $\alpha = 10$ and $\beta = -0.5$ and assumed a probit model.

277

When the subscription fee is \$22, our probit model would calculate the probability of a purchase as $\Phi(10 - 0.5 \times 22) = \Phi(-1) = 0.159$. We get the value of 0.159 using a standard spreadsheet formula for the normal distribution, as we cannot calculate it manually. For example, we can use NORM.S.DIST(-1,TRUE) in Excel to make this calculation for us. When we raise the price to \$23, the probability of a purchase falls to $\Phi(-1.5) = 0.067$. Hence, the effect of raising the subscription fee from \$22 to \$23 is to lower the probability of a purchase by 0.092 ($0.067 - 0.159 = -0.092$).

In contrast, suppose instead we knew $\alpha = 8$ and $\beta = -0.4$, and assumed a logit model. Consider now the same marginal effect calculation for a change in price from \$22 to \$23. When the subscription fee is \$22, our logit model would calculate the probability of a purchase as $\frac{1}{1+e^{-8+0.4\times22}} = \frac{1}{1+e^{0.8}} = 0.310$. When we raise the price to \$23, the probability of a purchase falls to $\frac{1}{1+e^{1.2}} = 0.231$. Thus, the effect of raising the subscription fee from \$22 to \$23 is to lower the probability of purchase by 0.079 ($0.231 - 0.310 = -0.079$).

We conclude by highlighting two additional features of marginal effects

for probit and logit models that lie in contrast to the linear probability model. The first is that probit and logit marginal effects generally depend on the magnitude of the change in the independent variable that we consider. To simplify intuition and exposition, our “rough” definition centers on one-unit increases in the X s. For the linear probability model, whose marginal effects are constant for a given X , this simplification is not particularly consequential. However, for probit and logit models, we may arrive at notably different calculations for a marginal effect when we consider the rate of change in $\Pr(Y = 1)$ for an increase in X by 1 versus, say, an increase in X by 0.1. Consequently, if we are interested in the effect of a fractional change in X (a marginal change that is notably less than one unit), we should adjust our formula for the marginal effect accordingly. Specifically, if we want to know the marginal effect of an increase in X by c , we can calculate it as:

$$\text{MargEff}_{X_j,C} = \frac{\Pr(Y_i = 1|X_{1i}, \dots, X_{ji} + C, \dots, X_K) - \Pr(Y_i = 1|X_1, \dots, X_{ji}, \dots, X_K)}{C}$$

To illustrate with our SaferContent example, suppose again that we knew $\alpha = 10$ and $\beta = -0.5$, and we assume a probit model. In addition, suppose we wanted to measure the marginal effect of a \$0.10 increase in price from our original price of \$22. As before, when the subscription fee is \$22, our probit model would calculate the probability of a purchase as $\Phi(10 - 0.5 \times 22) = \Phi(-1) = 0.159$. When we raise price to \$22.10, the probability of a purchase falls to $\Phi(-1.05) = 0.147$. The marginal effect of this \$0.10 price increase then is: $\frac{0.147 - 0.159}{0.1} = -0.12$. It is important to note that this measure is the *rate* of change in the probability of a subscription. If we wanted to predict the actual change in subscription probability when price increases from \$22 to \$22.10, we simply calculate the difference in the respective probabilities: $0.147 - 0.159 = -0.012$.

When computers automatically report marginal effects for continuous variables, they generally use the above formula for an infinitesimally small

value of c , applying some basics of calculus. While the full details of this calculation are beyond the scope of this book, note that we can calculate the rate of change in $\Pr(Y = 1)$ for any change, c , in X using the above formula. Our “rough” approach to measuring the rate of change in $\Pr(Y = 1)$, using one-unit increases in X , is appropriate for most applications.

278

The second feature of probit and logit model marginal effects that notably differs from the linear probability model is related to the first. Just as both models’ marginal effects can differ depending on the size of the change in X we are considering, their marginal effects also generally differ depending on the level of X from which we are considering a change. In our SaferContent example, this means that the marginal effect we would measure for a \$1 increase in price from \$22 to \$23 is generally *not* the same as the marginal effect we would measure for a \$1 increase in price from \$18 to \$19. To see this, consider once more the case where $\alpha = 10$ and $\beta = -0.5$, and we assume a probit model. We know from our earlier calculations that the marginal effect of increasing price from \$22 to \$23 is -0.092 . The marginal effect of increasing price from \$18 to \$19 is: $\Phi(0.5) - \Phi(1) = 0.691 - 0.841 = -0.15$. Notice that these marginal effects notably differ. In contrast, for the linear probability model, the marginal effect is constant and so will not depend on the price from which we make a change.

Because the marginal effects we measure depend on the starting point for X , there is not an obvious, single number to report as *the* marginal effect of X . Nevertheless, in practice, it is common to attempt to summarize the marginal effect of X for a probit or logit model using a single number. A typical way to do this is to calculate the marginal effect of X when starting from its mean value and setting all other X s to be their mean values. Within a general model, we would summarize the marginal effect of X_j by calculating:

$$\text{MargEff}_{X_j} = \Pr(Y_i = 1 | \bar{X}_{1i}, \dots, \bar{X}_{ji} + 1, \dots, \bar{X}_K) - \Pr(Y_i = 1 | \bar{X}_{1i})$$

Thus, as the average subscription price in our SaferContent example (from Table 9.2) is \$22.33, the single marginal effect we would calculate is:

$$\Pr(\text{Purchase}_i = 1 | \text{SubFee}_i = \$23.33) - \Pr(\text{Purchase}_i = 1 | \text{SubFee}_i =$$

ESTIMATION AND INTERPRETATION

LO 9.6 Execute estimation of a probit and logit model via maximum likelihood.

In our discussion of marginal effects for probit and logit models, we have taken the parameters (e.g., α, β) as given. In practice we must get estimates for these parameters using the data. For the linear probability model, this process is exactly the same as for our regression model—that is, solve for the parameters using the sample moment equations. In essence, solving the moment conditions for a regression model intends to make our predicted values for the outcome (i.e., $\widehat{Y}_i = \widehat{\alpha} + \widehat{\beta}_1 X_{1i} + \dots + \widehat{\beta}_K X_{Ki}$) best describe the actual outcomes (Y_i). However, our probit and logit models do not give predicted values for the outcomes but rather probabilities that the outcome equals 1. Hence, we cannot simply extend the use of sample moment equations to get estimates for the parameters of the determining function for our latent variable ($\alpha, \beta_1, \dots, \beta_K$).

Before laying out the details of an alternative approach to solving sample moment equations, we build some basic intuition. Suppose you are the coach of a youth basketball team, and a new player, Annie, has just joined the team. You want to get a sense of Annie's skill level before you start coaching her, and one skill you'd like to learn about is her ability to make free throws. To do this, you conjecture that as of joining the team, Annie's

knowledge of Annie, ρ could be anything between 0 and 1—she could miss everything, make everything, or anything in between.

How might you learn more about ρ ? The simplest and most intuitive way is to have Annie shoot free throws. Suppose she shoots ten free throws, and makes three, or 30% of her free throws—that is, she makes free throws at a rate of 0.3. A natural guess for Annie's likelihood of making any given free throw is $\rho = 0.3$.

Consider now an alternative calculation we could make using the data on Annie's ten free throws. We could ask, for a given probability of making a free throw (ρ), what is the probability that Annie makes three out of ten free throws—that is, what is the probability of observing the data we just collected? Such a calculation is straightforward. If we treat each shot as being independent, then the probability of a sequence of ten shots is just the product of the probability of each individual shot. Therefore, the probability of three makes and seven misses is: $\rho^3 \times (1 - \rho)^7$. In words, we have the product of three makes, each with probability ρ , times the product of seven misses, each with probability $(1 - \rho)$. Suppose we asked what value of ρ maximizes the probability of three makes and seven misses—what value of ρ makes a sequence of ten shots, with three makes and seven misses—most likely. It turns out that the solution to this problem yields our “natural” guess for ρ : 0.3!

We can even generalize the insight we just gained from Annie's free throws. Suppose we have a dichotomous variable, Y , that equals either 0 or 1. We observe N independent realizations of Y in a dataset, of which M realizations equal 1 (meaning $N - M$ realizations equal 0). Let ρ be the probability Y equals 1 for any given realization. Then, the value of ρ that maximizes $\rho^M \times (1 - \rho)^{(N-M)}$ is simply $\rho = M/N$. The value of ρ that maximizes the probability of what we observed is the rate at which Y equals 1 in the sample (M/N). Hence, the value for ρ that maximizes the likelihood of what we saw leads us to our natural guess for ρ , the number of times Y equals 1 divided by the number of observations of Y in the data. Linking this back to the free throw example, for Annie $N = 10$ and $M = 3$, and our solution for ρ is 0.3 (= 3/10).

The intuition underlying our example of Annie's free throws guides how we get our parameter estimates in probit and logit. The key difference for the probit and logit models is that they generally allow for different probabilities of Y equaling 1 for different values of X , whereas for Annie there was just a single probability, ρ . For given parameters of the determining function, the probit and logit models give us the probability of Y equaling one for each different possible value of X (according to the formulas in [Reasoning Box 9.2](#)). Consequently, solving for the parameters of the determining function using a dataset isn't as simple as calculating the rate at which Y equals 1 in the data—such a calculation yields just one number, and we need to solve for all the parameters of the determining function ($\alpha, \beta_1, \dots, \beta_K$).

Fortunately, we can utilize the alternative approach we introduced for Annie's free throws to get around this problem. We can ask what values for $(\alpha, \beta_1, \dots, \beta_K)$ within a probit or logit model make the observed values for Y (a set of 0s and 1s) as likely as possible. The solution, just as in our example with Annie, will give us a best guess for the true parameters of the determining function. This approach, where we estimate population-level parameters using values that make the observed outcomes as likely as possible for a given model, is known as **maximum likelihood estimation (MLE)**.

maximum likelihood estimation (MLE) Population-level parameters are estimated using values that make the observed outcomes as likely as possible for a given model.

280

To see how maximum likelihood estimation works for logit and probit models, consider once again a general latent variable model. Let

$$Y_i^* = \alpha + \beta_1 X_{1i} + \dots + \beta_K X_{Ki} + U_i$$

be the latent variable, and Y_i be 1 if the latent variable exceeds 0, and 0 if

otherwise. Suppose now we assume a probit model, so we have

$$\Pr(Y_i = 1 | X_{1i}, \dots, X_{Ki}) = \Phi(\alpha + \beta_1 X_{1i} + \dots + \beta_K X_{Ki})$$

To get estimates for our parameters, we collect a sample of Y s and X s of size N . Using maximum likelihood estimation, for probit we solve:

Max

$$(\alpha, \beta_1, \dots, \beta_k) \prod_{i=1}^N \Phi(\alpha + \beta_1 X_{1i} + \dots + \beta_K X_{Ki})^{Y_i} \times (1 - \Phi(\alpha + \beta_1 X_{1i} + \dots + \beta_K X_{Ki}))^{1 - Y_i}$$

The above expression may look complex, but it nicely lines up with the intuition from Annie's free throws. First, note that $\prod_{i=1}^N$ means we are multiplying N different expressions—one for each observation in our data. Note that if a given observation has $Y_i = 1$, the first half of the expression has an exponent of 1 and the second half has an exponent of 0 (thus making that term equal 1). This formulation captures that the probability of this observation is

$$\Phi(\alpha + \beta_1 X_{1i} + \dots + \beta_K X_{Ki})$$

that is, the probability that $Y_i = 1$, given the values of the corresponding X s. In contrast, if a given observation has $Y_i = 0$, the first half of the expression has an exponent of 0 (thus making that term equal 1) and the second half has an exponent of 1. Again, we see that this formulation captures that the probability of this observation is

$$1 - \Phi(\alpha + \beta_1 X_{1i} + \dots + \beta_K X_{Ki})$$

that is, the probability that $Y_i = 0$, given the values of the corresponding X s.

For logit, everything is exactly the same, except the probability formulas. Maximum likelihood estimation for logit solves:

Max

$$(\alpha, \beta_1, \dots, \beta_k) \prod_{i=1}^N \left(\frac{1}{1+e^{-\alpha-\beta_1 X_{1i}-\dots-\beta_K X_{Ki}}} \right)^{Y_i} \times \left(1 - \frac{1}{1+e^{-\alpha-\beta_1 X_{1i}-\dots-\beta_K X_K}} \right)^{1-Y_i}$$

Deriving the solution to these maximization problems requires a bit of calculus and is outside the scope of this book. Even if we derived the solution, it generally takes a computer to do the calculation. Fortunately, most statistical software will easily solve it with a single command (e.g., the commands “probit” or “logit” in STATA).

Just as with our regression model, we want to know that the MLE estimators we use for probit and logit models have “desirable” properties. We want to know if and when they are consistent and we can use them to build confidence intervals and conduct hypothesis tests. [Reasoning Boxes 9.3, 9.4, and 9.5](#) establish the basic conditions that lead to these properties. It is important to note that for simplicity some technical,

281

REASONING BOX 9.3

CONSISTENCY OF MLE ESTIMATORS FOR PROBIT/LOGIT DETERMINING FUNCTIONS

For a population of all possible realizations of Y, X_1, \dots, X_K , let $\{Y_i, X_{1i}, \dots, X_{Ki}\}_{i=1}^N$ be a sample of size N from this population, where

Y is a dichotomous variable. Further, let $(\hat{\alpha}, \hat{\beta}_1, \dots, \hat{\beta}_K)$ be the result of maximum likelihood estimation after assuming U_i conforms to the probit or logit model.

IF:

1. The data-generating process for an outcome, Y , can be expressed using a latent variable formulation with latent variable

$$Y_i^* = \alpha + \beta_1 X_{1i} + \dots + \beta_K X_{Ki} + U_i \text{ and:}$$

$$Y_1 = \begin{cases} 1 & \text{if } Y_i^* > 0 \\ 0 & \text{if } Y_i^* \leq 0 \end{cases}$$

2. $\{Y_i, X_{1i}, \dots, X_{Ki}\}_{i=1}^N$ is a random sample
3. U_i is independent of X_{1i}, \dots, X_{Ki}

THEN:

$(\hat{\alpha}, \hat{\beta}_1, \dots, \hat{\beta}_K)$ are consistent estimators of their corresponding parameters for the determining function, $(\alpha, \beta_1, \dots, \beta_K)$. We write this result as:

$$\begin{aligned} \hat{\alpha} &\rightarrow \alpha \\ \hat{\beta} &\rightarrow \beta_1 \\ &\dots \\ \hat{\beta} &\rightarrow \beta_K \end{aligned}$$

“regularity” conditions (which are nearly always met) have been omitted, and as with the regression model, we do not present the formulas for the standard errors of the estimators (e.g., $S_{\hat{\alpha}}$). In addition, note that we assume the

unobservables are *independent* of the X s. Recall from [Chapter 3](#) that this means the distribution of U is not affected by the value(s) of the X s. This assumption implies the unobservables and the X s are uncorrelated (thus mimicking the moment equations for the standard linear regression model). We must make this stronger assumption of independence for the probit and logit models because of their more complicated structure relative to the linear regression model (they are both nonlinear).

Just as with the regression model, building confidence intervals and conducting hypothesis tests for probit and logit models requires us to expand our assumptions beyond those needed to establish consistency. As before, we need a “big enough” sample. However, unlike for the regression model, we do not need to make an explicit assumption about homoscedasticity. This is because, for both the probit and logit, the corresponding assumption about the unobservables—that $U_i \sim N(0,1)$ or $U_i \sim \text{Logistic}(0,1)$ —already imposes homoscedasticity; both models have unobservables with constant variance. We need not restate this assumption.

282

REASONING BOX 9.4

CONFIDENCE INTERVALS FOR PARAMETERS OF A PROBIT/LOGIT DETERMINING FUNCTION

For a population of all possible realizations of Y, X_1, \dots, X_K , let $\{Y_i, X_1, \dots, X_{Ki}\}_{i=1}^N$ be a sample of size N from this population, where Y is a dichotomous variable. Further, let $(\hat{\alpha}, \hat{\beta}_1, \dots, \hat{\beta}_K)$ be the result of maximum likelihood estimation after assuming U_i conforms to the probit or logit model.

IF:

1. The data-generating process for an outcome, Y , can be expressed using a

latent variable formulation with latent variable

$$Y_i^* = \alpha + \beta_1 X_{1i} + \dots + \beta_K X_{Ki} + U_i \text{ and:}$$

$$Y_i = \begin{cases} 1 & \text{if } Y_i^* > 0 \\ 0 & \text{if } Y_i^* \leq 0 \end{cases}$$

2. $\{Y_i, X_{1i}, \dots, X_{Ki}\}_{i=1}^N$ is a random sample
3. U_i is independent of X_{1i}, \dots, X_{Ki}
4. The size of the sample is at least $30 \times (K + 1)$

THEN:

The interval consisting of $\hat{\alpha}$ plus or minus 1.65 (1.96, 2.58) times $S_{\hat{\alpha}}$ will contain α approximately 90% (95%, 99%) of the time. The same holds true for $\hat{\beta}_1, \dots, \hat{\beta}_K$.

Inductive reasoning: Based on the observation of $\hat{\alpha}, S_{\hat{\alpha}}$, and N , α is contained in the interval $(\hat{\alpha} \pm 1.65 (S_{\hat{\alpha}}))$. The objective degree of support for this inductive argument is 90%. If we instead use the intervals $(\hat{\alpha} \pm 1.96 (S_{\hat{\alpha}}))$ and $(\hat{\alpha} \pm 2.58 (S_{\hat{\alpha}}))$, the objective degree of support becomes 95% and 99%, respectively.

The same holds true for $\hat{\beta}_1, \dots, \hat{\beta}_K$.

REASONING BOX 9.5

HYPOTHESIS TESTING FOR PARAMETERS OF A PROBIT/LOGIT DETERMINING FUNCTION

For a population of all possible realizations of Y, X_1, \dots, X_K , let

$\{Y_i, X_{1i}, \dots, X_{Ki}\}_{i=1}^N$ be a sample of size N from this population, where Y is a dichotomous variable. Further, let $(\hat{\alpha}, \hat{\beta}_1, \dots, \hat{\beta}_K)$ be the result of

maximum likelihood estimation after assuming U_i conforms to the probit or logit model.

IF:

1. The data-generating process for an outcome, Y , can be expressed using a latent variable formulation with latent variable

$$Y_i^* = \alpha + \beta_1 X_{1i} + \dots + \beta_K X_{Ki} + U_i \text{ and:}$$

$$Y_i = \begin{cases} 1 & \text{if } Y_i^* > 0 \\ 0 & \text{if } Y_i^* \leq 0 \end{cases}$$

283

2. $\{Y_i, X_{1i}, \dots, X_{Ki}\}_{i=1}^N$ is a random sample
3. U_i is independent of X_{1i}, \dots, X_{Ki}
4. The size of the sample is at least $30 \times (K + 1)$
5. $\alpha = c_0$

THEN:

We have $\hat{\alpha} \sim N(c_0, \sigma_\alpha)$ and $\hat{\alpha}$ will fall within 1.65 (1.96, 2.58) standard deviations of c_0 approximately 90% (95%, 99%) of the time. This also means that $\hat{\alpha}$ will differ by more than 1.65 (1.96, 2.58) standard deviations from c_0 (in absolute value) approximately 10% (5%, 1%) of the time.

The same holds true for each of $\hat{\beta}_1, \dots, \hat{\beta}_K$ when assuming, e.g., $\beta_j = c_j$

Inductive reasoning:

Using t-stats. If the absolute value of the t-stat for $\hat{\alpha}$ ($= \left| \frac{\hat{\alpha} - c_0}{S_{\hat{\alpha}}} \right|$) is greater than 1.65 (1.96, 2.58), reject the deduced (above) distribution for $\hat{\alpha}$. Otherwise, fail to reject. The objective degree of support for this inductive argument is 90% (95%, 99%).

The same holds true for $\hat{\beta}_1, \dots, \hat{\beta}_K$ when assuming, e.g., $\beta_j = c_j$.

Using p-values. If the p -value of the t -stat for $\hat{\alpha}$ is less than 0.10 (0.05, 0.01), reject the deduced (above) distribution for $\hat{\alpha}$. Otherwise, fail to reject. The objective degree of support for this inductive argument is 90% (95%, 99%).

The same holds true for $\hat{\beta}_1, \dots, \hat{\beta}_K$ when assuming, e.g., $\beta_j = c_j$.

Transposition:

If inductive reasoning leads to a rejection of the distribution for $\hat{\alpha}$, reject at least one of the assumptions (1, 2, 3, 4, or 5) leading to that distribution. If there is confidence in assumptions 1–4, this means rejection of the null hypothesis.

Now that we have stated the general properties of MLE estimators for probit and logit models, we conclude by showing probit and logit estimations for our SaferContent example. Consider again our original data for the SaferContent example from [Table 9.3](#). Suppose we want to estimate the relationship between subscription fee and the probability of a purchase using a probit or logit model, rather than a linear probability model.

For both the probit and logit models, we assume $\text{Utility}_i = \alpha + \beta \text{SubFee}_i + U_i$. The data-generating process is such that:

$$\text{Purchase}_i = \begin{cases} 1 & \text{if } \alpha + \beta \text{SubFee}_i + U_i > 0 \\ 0 & \text{if } \alpha + \beta \text{SubFee}_i + U_i \leq 0 \end{cases}$$

For probit, we assume $U_i \sim N(0,1)$. With this assumption, we can estimate α and β using maximum likelihood. We present the results in [Table 9.8](#) (calculated using, e.g., STATA's "probit" command).

From these estimates, we see that an increase in the subscription fee of \$1 reduces Utility by 0.146, on average. Further, from the p -value, we see that we reject the hypothesis that $\beta = 0$ with very high confidence, and from the confidence interval, we see that we

TABLE 9.8 Probit Results for SaferContent Data

	COEFFICIENTS	STANDARD ERROR	Z SCORE ¹	P-VALUE	LOWER 95%	4
Intercept	3.204304	0.901809	3.55	0.000	1.436791	4
SubFee	-0.1463224	0.0400753	-3.65	0.000	-0.2248685	-0

¹Note that these tables use a z score rather than a t-statistic. This distinction is meaningful for small samples, but since we are assuming sufficiently large sample sizes, the difference is merely semantic, as the p-values are calculated in the same way.

are 95% confident that an increase in subscription fee will reduce Utility by somewhere between 0.068 and 0.225.

Of course, changes in Utility are not generally our primary interest. Rather, we want to know how a change in subscription fee will affect purchase decisions via the probability of a purchase. That is, we want to know the marginal effect for the subscription fee. If there is not a specific subscription fee for which we want to know the marginal effect, then we can calculate it for the average subscription fee, as is typical. In our sample, the average subscription fee is \$22.33, so we can calculate the marginal effect as:

$$\Phi(3.204 - 0.146 \times 23.33) - \Phi(3.204 - 0.146 \times 22.33) = 0.420 - 0.478 = -0.058$$

Thus, a simple summary of the effect of raising subscription fee by \$1 is that it lowers the probability of purchase by nearly 6%. Note, though, that this relationship likely will differ if we start at a different subscription fee (e.g., \$25).

If we instead use the logit model, we assume $U_i \sim \text{Logistic}(0,1)$. With this assumption, we again can estimate α and β using maximum likelihood. We present the results in [Table 9.9](#) (calculated using, e.g., STATA's "logit" command).

From these estimates, we see that an increase in the subscription fee of \$1 reduces Utility by 0.242, on average. Further, from the *p*-value, we see that we reject the hypothesis that $\beta = 0$ with very high confidence. Also, from the confidence interval, we see that we are 95% confident that an increase in subscription fee will reduce Utility by somewhere between 0.105 and 0.379. The marginal effect is:

$$\frac{1}{1 + e^{-5.295+0.242 \times 23.33}} - \frac{1}{1 + e^{-5.295+0.242 \times 22.33}} = 0.413 - 0.473 = -0.060$$

Thus, a simple summary of the effect of raising subscription fee by \$1 is that it lowers the probability of purchase by about 6%. As with probit, this relationship likely will differ if we start at a different subscription fee (e.g., \$25). However, notice that we arrived at very similar marginal effects for probit and logit, reflecting the general similarities in the two models.

TABLE 9.9 Logit Results for SaferContent Data

	COEFFICIENTS	STANDARD ERROR	Z SCORE	P-VALUE	LOWER 95%	UPP 95
Intercept	5.29481	1.566023	3.38	0.001	2.225461	8.364
SubFee	-0.2417501	0.0698248	-3.46	0.001	-0.3786043	-0.10

9.2

Demonstration Problem

An outcome variable, Y , can only take on the values zero or one. In attempting to measure the effects of X_1 and X_2 on Y , you've collected a sample of size 200 on these three variables, and assumed the following:

- A. The data generating process for Y is:

$$Y_i = \begin{cases} 1 & \text{if } Y_i^* > 0 \\ 0 & \text{if } Y_i^* \leq 0 \end{cases}$$

where $Y_i^* = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + U_i$

- B. $\{Y_i, X_{1i}, X_{2i}\}_{i=1}^N$ is a random sample
- C. U_i is independent of X_{1i}, X_{2i}

You further assume a probit model, and get the following parameter estimates using maximum likelihood estimation (e.g., using the command “probit” in STATA):

TABLE 9.10 Probit Results for Data on Y , X_1 , and X_2

	COEFFICIENTS	STANDARD ERROR	Z SCORE	P-VALUE	LOWER 95%	UPPER 95%
Intercept	-9.7721	1.2720	-7.68	0.000	-12.2652	-7.279
X1	0.2493	0.0360	6.93	0.000	0.1788	0.3198
X2	0.7269	0.1019	7.14	0.000	0.5273	0.9266

- a. Interpret the estimates for β_1 and β_2 .
- b. What is the marginal effect of X_1 when $X_1 = 5$ and $X_2 = 9$?
- c. What is the marginal effect of X_2 when $X_1 = 5$ and $X_2 = 9$?
- d. Are there any values for X_1 and X_2 that could generate a limit-violating prediction?

Answer:

- a. The estimate for β_1 implies that, when X_1 increases by one unit and X_2 is held constant, Y^* increases by 0.2493. The estimate for β_2 implies that, when X_2 increases by one unit and X_1 is held constant, Y^* increases by 0.7269. If, for example, Y^* represents utility, these changes are in utils. However, we are seldom interested in the changes to Y^* in a probit (or logit) model.
- b. The marginal effect at these X_1 and X_2 values is: $\Phi(-9.7721 + 0.2493 \times 6$

$$+ 0.7269 \times 9) - \Phi(-9.7721 + 0.2493 \times 5 + 0.7269 \times 9) = 0.0178.$$

- c. The marginal effect at these X_1 and X_2 values is: $\Phi(-9.7721 + 0.2493 \times 5 + 0.7269 \times 10) - \Phi(-9.7721 + 0.2493 \times 5 + 0.7269 \times 9) = 0.0808$.
- d. No. This is one of the merits of the probit (and logit) model. By construction, probabilities must always be between 0 and 1.

MERITS AND SHORTCOMINGS

LO 9.7 Identify the merits and shortcomings of probit and logit models in practice.

We conclude our discussion of probit and logit models with a brief summary of their merits and shortcomings. The key merits link back to the linear probability model, in that probit and logit models help to overcome shortcomings of the linear probability model. Recall that the linear probability model is such that: (1) The determining function and

286

unobservables always add up to exactly 0 or 1, and (2) limit-violating predictions are possible.

Probit and logit models overcome these shortcomings by construction. As noted above, the latent variable formulation places no restrictions per se on the relationship between the determining function and unobservables, thus avoiding shortcoming (1). Further, both models predict probabilities rather than the actual value (0 or 1) for the dependent variable. Since the prediction is already a probability and forced to be between 0 and 1 by construction, there is no risk of limit-violating predictions when considering a change in an independent variable. Hence, both models avoid shortcoming (2).

While probit and logit models do overcome some of the key shortcomings of the linear probability model, they aren't without their own shortcomings. To conclude, we highlight some notable shortcomings of these models.

First, note that the probabilities implied by the probit and logit models directly depend on the assumption of a normal or logistic distribution for the unobservables, respectively. Hence, our estimates for the parameters of the determining function, and also the marginal effects, will be inconsistent if this assumption is incorrect, no matter the sample size. In contrast, estimates for the linear probability model are not materially affected by the distribution of the unobservables, as long as the sample is large. While this criticism may be important in some special cases, the assumption of normally or logically distributed unobservables (which are very similar in shape) is often a reasonable one, and so this shortcoming often is not especially problematic in practice.

A second shortcoming of probit and logit models is the added complexity of calculating marginal effects, relative to the linear probability model. Recall that marginal effects for probit and logit depend on the levels of all of the X s, whereas marginal effects for the linear probability model are constant. Of course, the linear probability model may be oversimplifying these effects, but exposition and interpretation are substantially simpler.

A final shortcoming of probit and logit models concerns the use of instrumental variables and fixed effects. Just as in the standard regression model, we may be worried about endogeneity within a probit or logit model. And we may want to utilize an instrumental variable(s) and/or fixed effects to address this issue. Unfortunately, utilizing these methods within a probit or logit model is a bit more complicated than it is for a linear probability model, and typically requires additional assumptions. The details of how to employ instrumental variables and fixed effects using probit and logit are beyond the scope of this book.

COMMUNICATING DATA 9.3

THE “RIGHT” MODEL FOR A DICHOTOMOUS DEPENDENT VARIABLE

Throughout the text of this chapter, notice we give no indication of which model (linear probability, probit, logit) is “right,” but just their merits and shortcomings. We conclude this chapter using a Communicating Data discussion in order to finish with more global points concerning model selection for a dichotomous dependent variable. First, note that there is little practical difference between the probit and logit models, so the key choice is between probit/logit and the linear probability model. Second, we are generally interested in marginal effects, so differences in estimated marginal effects will be the key distinctions across the models. Third, the models will give, by construction, different estimates for marginal effects across

287

different starting values for the X s. However, it is not unusual for data to be concentrated around a relatively small range of X values with a relatively small range of corresponding probabilities for Y . In such cases, the estimated marginal effects for the two different types of models generally won’t be very different when starting from X values in the observed range, and so the choice between models is not particularly crucial.

For cases where the data are more expansive and/or we want marginal effects for notably varying X levels, the differences in estimated marginal effects across the models can be large. For these cases, there is no consensus on the “right” model to use. Hashing out the full set of arguments in either direction is too large an issue to fully tackle here, but we conclude by noting both rely on assumptions that are difficult to test. Thus, predictions in such cases should recognize this fact and carry appropriate caveats.

RISING TO THE dataCHALLENGE

Changing the Offer to Change Your Odds

The decision to accept or decline an offer is a dichotomous one, so we can relabel Accept as 1 and Decline as 0. We would like to know the effect of raising salary by \$5,000 on the likelihood of acceptance. After working through

this chapter, we know the most commonly utilized models for measuring such an effect are the linear probability model and the probit/logit models. Suppose we choose to assume a logit model. Hence, we assume the data-generating process for Y is:

$$Y_i = \begin{cases} 1 & \text{if } \alpha + \beta \text{Offer}_i + U_i > 0 \\ 0 & \text{if } \alpha + \beta \text{Offer}_i + U_i \leq 0 \end{cases}$$

where $U_i \sim \text{Logistic}(0,1)$.

Using maximum likelihood estimation (e.g., “logit” in STATA), we get the estimates for α and β shown in Table 9.11.

TABLE 9.11 Logit Results for Offer and Acceptance Employment Data

	COEFFICIENTS	STANDARD ERROR	Z SCORE	P-VALUE	LOWER 95%	UPPER 95%
Intercept	-2.3373	1.9751	-1.18	0.237	-6.2085	1.5338
Offer (in '000s)	0.0324	0.0270	1.20	0.230	-0.0205	0.0852

Given we are using the logit model, measuring the effect of a \$5,000 increase in the offer depends on the starting point. For example, we may want to know the effect of a \$5,000 increase in the offer from a base level of \$60,000. Using our estimates, we have the measured effect:

$$\frac{1}{1 + e^{2.3373 - 0.0324 \times 65}} - \frac{1}{1 + e^{2.3373 - 0.0324 \times 60}} = 0.0395$$

From this calculation, our model implies that raising an offer from \$60,000 to \$65,000 should, on average, raise the likelihood of an acceptance by about 4%.

However, there are important caveats to this conclusion. First, the

coefficient on the offer was not statistically significant. Consequently, we cannot reject the hypothesis that acceptance was unaffected by the offer amount, and so our measured effect of a \$5,000 increase in the offer—while a best guess—may be an overstatement. However, it seems intuitively hard to believe that offering more money would not raise the likelihood of acceptance.

Perhaps the relatively small size of the dataset simply led to a particularly uncertain estimate for β , but we should be wary of another potential issue. In particular, we might worry that offers are endogenous within this model. That is, it may be the case that candidates with strong alternative employment options were given larger offers. Within the model, this means those with low values for U (unobservables making it less likely they will accept) were given high values for Offer. This clearly violates the assumption of U being independent of the Offer, meaning we cannot count on getting a consistent estimate for β . In this case, we likely are underestimating the effect of the Offer on Acceptance. Addressing this latter issue generally will require some of the remedies we've described in prior chapters, e.g., more controls, fixed effects, and/or the use of an instrumental variable.

SUMMARY

In this chapter we defined a limited dependent variable. We focused our attention on a specific type of limited dependent variable, the dichotomous dependent variable, which takes on just two values. We presented the linear probability model and explained how it applies regression analysis to dichotomous dependent variables to predict outcome probabilities. We then detailed important merits and shortcomings of the linear probability model.

We next described probit and logit models, which build upon latent variables, as alternatives to the linear probability model. We explained how to calculate changes in outcome probabilities, called marginal effects, using these more complex models. We went on to detail how to estimate a probit and logit model and interpret the results from each. We concluded by listing the merits and shortcomings of the probit and logit, and contrasting them with the linear probability model.

KEY TERMS AND CONCEPTS

dichotomous dependent variable

latent variable

limit-violating prediction

limited dependent variable

linear probability model

logit model

marginal effect

maximum likelihood estimation (MLE)

probit model

289

CONCEPTUAL QUESTIONS connect

1. Which of the following could reasonably be categorized as a limited dependent variable? (LO1)
 - a. Firm revenue
 - b. Number of automobiles owned by a household
 - c. Number of firm employees
 - d. Whether an individual is unemployed
 - e. Whether an employee was promoted last year
2. Which of the following could be categorized as a dichotomous dependent variable? (LO1)
 - a. Firm revenue
 - b. Whether an individual is unemployed
 - c. Number of automobiles owned by a household
 - d. If a household viewed a new shoe ad
 - e. Firm profits
3. Suppose Y is a dichotomous dependent variable, and the data-generating process for Y can be expressed as:

$$Y_i = 0.1 + 0.03X_{1i} - 0.02X_{2i} + U_i$$

Interpret the coefficients on X_1 and X_2 . (LO2)

4. Suppose you are trying to learn the relationship between the price you charge for your product and the likelihood of purchase by individuals offered that price. You offer your product online and for one month have randomly posted prices between \$10 and \$30. Using data on purchases and prices, you get the following estimates for a linear probability model:

$$\text{Purchase}_i = 1.7 - 0.06 \times \text{Price}_i$$

You are interested in the effect of a \$20 price increase (i.e., moving from the lowest price to the highest) on the likelihood of Purchase. Why is answering this question problematic using this model? (LO3)

5. You are interested in learning how the size of your website ad affects the likelihood of visitors to that website clicking the ad. Express the relationship between the incidence of a click and the size of the ad using a latent variable formulation. (LO4)
6. Suppose a latent variable, Y^* , can be expressed as $Y_i^* = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + U_i$. Suppose also that Y equals 1 when $Y^* > 0$ and Y equals 0 otherwise. (LO4)
- What other features must this data-generating process possess for us to label it as:
 - A logit model?
 - A probit model?
 - Name a reason for:
 - Assuming a logit model instead of a probit model
 - Assuming a probit model instead of a logit model
7. Which of the following are shortcomings of the linear probability model? (LO3)
- They can generate limit-violating predictions

- b. Marginal effects depend on the starting point
- c. They can rely on often unrealistic distributions for the unobservables
- d. Use of instrumental variables requires additional assumptions that generally are not needed for OLS

290

-
- 8. Which of the following are shortcomings of probit and logit models? (LO7)
 - a. They can generate limit-violating predictions
 - b. Marginal effects depend on the starting point
 - c. They can rely on often unrealistic distributions for the unobservables
 - d. Use of instrumental variables requires additional assumptions that generally are not needed for OLS
 - 9. Assume a latent variable, Y^* , can be expressed as $Y_i^* = 4 + 6X_i + U_i$, where U_i is independent of X_i . Suppose also that Y equals 1 when $Y^* > 0$ and Y equals 0 otherwise. Calculate the marginal effect of X when:
 - a. $X = -0.5$ and we assume this is a logit model.
 - b. $X = -1.2$ and we assume this is a probit model.
 - c. $X = -0.2$ and we assume this is a logit model.
 - 10. Assume a latent variable, Y^* , can be expressed as:

$$Y_i^* = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + U_i$$

where U_i is independent of X_{1i} and X_{2i} .

Suppose also that Y equals 1 when $Y^* > 0$ and Y equals 0 otherwise. Using a random sample of $\{Y_i, X_{1i}, X_{2i}\}$, you get the following maximum likelihood estimates: (LO5)

$$\hat{\alpha} = 2.3$$

$$\hat{\beta} = 1.2$$

$$\hat{\beta}_1 = -0.4$$

- a. What is the marginal effect of X_1 when $X_1 = 0$ and $X_2 = 5$ if we assume this is a *probit* model?
 - b. What is the marginal effect of X_2 when $X_1 = 0.2$ and $X_2 = 6$ if we assume this is a *logit* model?
11. Refer to Question 10. Suppose the actual distribution for U can be expressed as follows: (LO7)
-

$$\begin{aligned}\Pr(U_i = -2) &= 0.3 \\ \Pr(U_i = 0) &= 0.4 \\ \Pr(U_i = 2) &= 0.3\end{aligned}$$

- a. Why should you be concerned about using your above estimates if you assumed a probit model to get them?
- b. Would this concern diminish if you had assumed a logit model instead?
- c. Would this concern diminish if you tripled the sample size?

QUANTITATIVE PROBLEMS connect

Dataset available at www.mhhe.com/prince1e

12. You have begun consulting for Yummy Yogurt, and they are interested in learning the impact of a digital discount campaign the company recently began. The campaign consists of sending a price discount to prior customers on an infrequent basis—some weeks customers receive the discount while in other weeks they do not. Yummy Yogurt has collected data on the incidence of receiving the discount and whether a purchase was made for 1,000 customers over 16 weeks. The data are in *Chap9Prob1213*, and the unit of observation is an individual-week. Here, Purchase equals 1 if a purchase of Yummy Yogurt

291

was made in a given week, and 0 otherwise; Discount equals 1 if a discount was offered in a given week, and 0 otherwise.

Assume the following data-generating process: $\text{Purchase}_{it} = \alpha + \beta \text{Discount}_{it} + U_{it}$. Hence, you are assuming a linear probability model. Also, assume these data are a random sample and there is no correlation between unobservables and Discount. (LO2)

- a. Using OLS, get estimates for α and β .
 - b. Interpret your estimate for β .
 - c. At a 95% confidence level, can you reject the hypothesis that the discount has no effect on the likelihood of purchase?
 - d. Is there reason for concern about limit-violating predictions for this linear probability model?
- 13.** Refer to Problem 12. (LO2)

Dataset available at www.mhhe.com/prince1e

- a. What type of data are in *Chap9Prob1213*?
 - b. Why might it be important to include fixed effects in your estimation?
 - c. Write out the data-generating process for these data with fixed effects included.
 - d. Use a within estimator to get estimates for α and β .
 - e. Interpret your estimate for β .
- 14.** Your firm has begun advertising on YouTube, placing video ads at the beginning of popular music videos. You have many different versions of your ad, ranging from 15 seconds to 2 minutes in length. YouTube varies the length of the ads presented to viewers of its music videos, and you have data detailing which ad was viewed and whether that individual ultimately visited your website during that session online. The data are in *Chap9Prob1415*. Here, Visit equals 1 if your website was visited during an online session, and Ad Length is the length, in seconds, of the ad that individual viewed from your firm. As can be seen in the dataset, you also have information on the individual's age. You would like to learn the effect of ad length on the likelihood of visiting your website. You start by assuming there is a latent variable: (LO5)

Dataset available at www.mhhe.com/prince1e

$$Y_i^* = \alpha + \beta_1 \text{Ad Length}_i + \beta_2 \text{Age}_i + U_i$$

- a. Suppose you want to estimate a probit model. Detail the assumptions you must make, in addition to the latent variable definition above, in order to estimate the effect of ad length on the likelihood of a visit.
 - b. Assume a probit model, and use maximum likelihood estimation to estimate α , β_1 , and β_2 .
 - c. Using your estimates from Part b, what is the effect on the likelihood of a visit to your website from increasing ad length from 30 seconds to one minute for an individual who is 25 years old?
15. Refer to Problem 14. (LO6)

Dataset available at www.mhhe.com/prince1e

- a. Suppose you want to estimate a logit model instead. How do your assumptions differ from those in Part a of Problem 14 if you again want to estimate the effect of ad length on the likelihood of a visit?
- b. Assume a logit model, and use maximum likelihood estimation to estimate α , β_1 , and β_2 .
- c. Using your estimates from Part b, what is the effect on the likelihood of a visit to your website from increasing ad length from 30 seconds to one minute for an individual who is 25 years old?
- d. Suppose you believe the effect of ad length on the latent variable is not constant—for example, the effect of increasing ad length from 30 seconds to 45 seconds is not the same as increasing ad length from 1:30 to 1:45. How could you alter the determining function for Y^* to allow these effects to differ?

Identification and Data Assessment

10

LEARNING OBJECTIVES

After completing this chapter, you will be able to:

- LO10.1** Explain what it means for a variable's effect to be identified in a model.
- LO10.2** Explain extrapolation and interpolation and how each inherently suffers from an identification problem.
- LO10.3** Distinguish between functional form assumptions and enhanced data coverage as remedies for identification problems stemming from extrapolation and interpolation.
- LO10.4** Differentiate between endogeneity and types of multicollinearity as identification problems due to variable co-movement.
- LO10.5** Articulate remedies for identification problems and inference challenges due to variable co-movement.
- LO10.6** Solve for the direction of bias in cases of variable co-movement.

Chapter opener image credit: ©naqiewei/Getty Images

dataCHALLENGE Are Projected Profits over the Hill?

As an analyst for BabyWear—an online vendor of diapers—you have been given a rare opportunity to experiment with online prices in order to determine the relationship between profits and price. Over a large number of zip codes, you randomly vary the price of BabyWear’s standard package of newborn diapers between \$34.99 and \$38.99 and maintain these random prices for one month. You then record variable profits (equal to revenues minus variable costs) at each randomly chosen price. To conduct your analysis of the relationship between Profits and Price, you assume the following data-generating process:

$$\text{Profits}_i = \alpha + \beta \text{Price}_i + U_i$$

You then regress Profits on Price and get the estimates reported in, [Table 10.1](#).

TABLE 10.1 Regression Results for Regression of BabyWear Profits on Pricetabmar

	COEFFICIENTS	STANDARD ERROR	t STAT	P-VALUE
Intercept	716.1734129	22730.01617	0.031507827	0.974901295
Price	1503.742314	616.4179959	2.439484772	0.015729687

293

Given the randomness of your price assignment, you believe the relationship from your regression is causal, and you present it as such to BabyWear management. Based on these findings, management concludes profits would be decidedly higher if the company raised its usual price of \$36.99 to \$49.99.

Would you endorse this conclusion?

Introduction

In the prior chapters, a key point of emphasis has been the consequences of, and remedies for, endogeneity problems. When the treatment variable is endogenous within our assumed data-generating process, we are unable to consistently estimate its effect on the outcome. However, endogeneity is not the only reason we may be unable to estimate a treatment effect. In this chapter, we “widen the lens” with regard to challenges toward estimating a treatment effect. We discuss the range of dataset characteristics, including those leading to endogeneity, that would preclude us from estimating a treatment effect, regardless of the size of the dataset. We then detail remedies for such data limitations in the form of assumptions and/or changes to the data. We also discuss ways to bound the direction and magnitude of the inaccuracy in treatment-effect estimation when remedies are not sufficient or possible.

A broad understanding of data features that allows us to estimate a treatment effect in general is crucially important. It aids two audiences: It helps the analyst in collecting proper data and making appropriate assumptions, and it also helps consumers of the analysis assess whether the data being used are capable of estimating the treatment effect in which they are interested.

Assessing Data Via Identification

LO 10.1 Explain what it means for a variable’s effect to be identified in a model.

Suppose you are working for a craft furniture designer who has his own website. He is interested in how sensitive his consumers are to price for his hand-crafted rocking chairs. He charges the same price to everyone for the chairs themselves (\$200), but shipping costs from his production facility in central Wisconsin vary depending on the location of the customer. Thus, the

total price to customers varies by customer location. Shipping costs make a discrete jump when the customer is outside Wisconsin. They are between \$10 and \$25 in-state and between \$75 and \$100 out-of-state. The designer provides you with a cross-sectional dataset with observations that vary by location. In particular, the data contain information on locations (zip codes), the full price charged for a chair to be delivered to that location, and the total sales for that location over a one-year period. The first few observations are presented in [Table 10.2](#).

Your goal is to estimate the average treatment effect of price on sales. On average, when price increases by, say, \$1, what is the effect on sales of rocking chairs? As we have

294

TABLE 10.2 Subsample of Rocking Chair Data

IP LOCATION ZIP CODE	PRICE	SALES
90006	\$297.32	8
32042	283.75	9
45233	275.35	7
07018	280.25	11
53082	223.50	17
53039	214.10	12
37055	292.90	8

indicated in prior chapters, a natural concern when trying to answer this question is that price is correlated with other (unobserved) factors that influence sales, creating an endogeneity problem. However, before attempting to address this specific concern, we can ask whether a broader problem exists with the data we have—whether the effect of price on sales is identified.

In statistics, the definition of “identified” can be rather formal and complex. Here, we take a less formal, practically oriented, approach toward

defining the concept of “identified”—an approach that is tailored to the idea of estimating causal effects of variables on outcomes. In our analysis of sales and price for rocking chairs, suppose we assumed the following data-generating process:

$$\text{Sales}_i = \alpha + \beta \text{Price}_i + U_i$$

Within this model, we are interested in accurately estimating β . We say that a parameter (e.g., β) is **identified** within a given model if it can be estimated with any level of precision given a large enough sample from the population. A bit more formally, a parameter is identified if, for a given confidence level K ($< 100\%$) and a given “length” L , we can build a confidence interval that contains β with length less than L and confidence level of K , given a large enough sample of data. For example, suppose we want to estimate a parameter with a confidence level of 99% and a confidence interval whose upper and lower bounds differ by only 0.1. If that parameter is identified, we can generate such a confidence interval given enough data.

identified Can be estimated with any level of precision given a large enough sample from the population.

While our ultimate goal is to assess when and how parameters of an assumed data-generating process (like the one assumed above) are identified, it is best to start within a simpler framework to build our understanding of identification. Consider a simple die—a cube with the numbers 1 through 6 on its faces and with exactly one number on each face. Suppose we want to know the probability of rolling the number 3 for that die. If it is a “fair” die, that probability is $1/6$. However, suppose we aren’t sure it is a fair die—perhaps it is lopsided in some way affecting the probability each number is rolled (a trick often used by cheats in the game of craps at casinos). We’d like to be able to determine what the “true” probability of rolling a 3 is for that die with a high level of precision. How could we do it? An attractive option is to do it empirically, using data on rolls for that die.

To begin, define p as the probability of rolling a 3 on any single roll of

the die. Then, if we define X to be the number of 3s we observe on a single roll of the die ($X = 1$ if we roll a 3 and $X = 0$ if we roll any other number), we have $E[X] = p$. Some additional math would

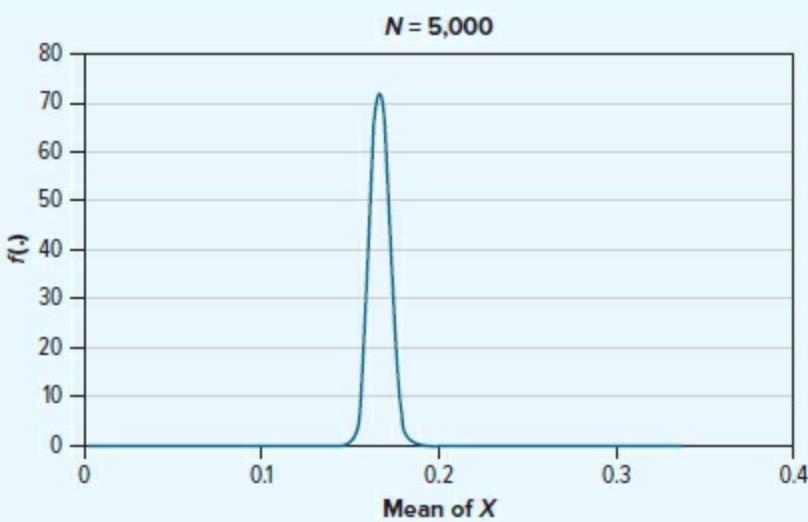
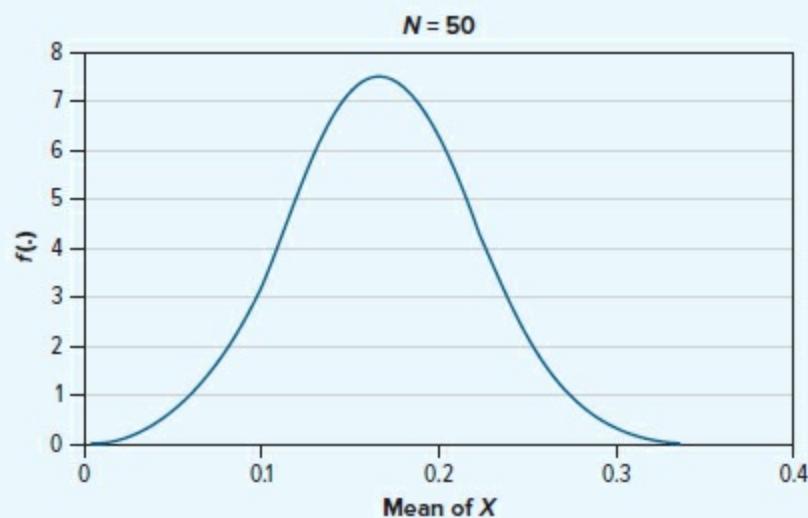
295

show that $\text{Var}[X] = p(1 - p)$. Within this simple framework, the parameter p is identified. In other words, we can estimate p as precisely as we want given enough data on the die (given enough rolls of the die).

The fact that p is identified follows directly from the central limit theorem (discussed in [Chapter 3](#)). Suppose we roll the die N times. Define x_1 as the observed value of X for the first roll, x_2 for the second, and so on. Then, define $\bar{X}_N = \frac{1}{N}[x_1 + x_2 + \dots + x_N] = \frac{1}{N} \sum_{i=1}^N x_i$, i.e., the sample mean for X , or equivalently, the proportion of the N rolls that showed a 3. Given these definitions, the central limit theorem states that $\bar{X}_N \sim N\left(p, \frac{\sqrt{p(1-p)}}{\sqrt{N}}\right)$ as N gets large.

Rather than dive into a formal proof from here, consider what this statement about the distribution of \bar{X}_N implies both visually and conceptually. Visually, we present two distributions for \bar{X}_N in [Figure 10.1](#)—both have $p = 1/6$, but one has $N = 50$ and the other has $N = 5,000$. We can see that the distribution “collapses” around the value for p . As N gets

FIGURE 10.1 Distribution of Mean of X for $N = 50$ and $N = 5,000$



296

REASONING BOX 10.1

CAN DATA DELIVER THE (SUFFICIENTLY PRECISE) ANSWER?

Suppose we have determined a population of data from which we are able to sample, and a treatment effect we would like to estimate. If there is an

acceptable model of the data-generating process (one whose accompanying assumptions are believable) within which the treatment effect is identified using samples from the population, then the treatment effect can be estimated with any level of precision given a large enough sample of these data. Under these circumstances, an analyst can estimate the desired treatment effect without seeking alternative data populations.

In contrast, if there is *not* an acceptable model of the data-generating process within which the treatment effect is identified using samples from the population, the analyst should consider alternative data populations (e.g., additional variables) before attempting to estimate the effect.

larger, any realized value for \bar{X}_N is almost certainly very close to p ($= 1/6$). Put another way, as we build confidence intervals for p using these samples, the confidence intervals tend to get smaller and smaller for the same level of confidence. Conceptually, the idea is quite simple: the mean of a sample will not stray far from the population mean as the sample keeps getting larger. More data—sampled randomly—will get us an estimate as close to p as we want.

When considering questions concerning the effect of a strategic variable on a particular outcome, we want to be sure the data we've collected are *capable* of providing a useful answer. For this to be true, we must be able to find an acceptable model of the data-generating process (whose accompanying assumptions are believable) within which the effect we are trying to estimate is identified using the population of data to which we have access. If so, our only remaining concern is the level of precision we hope to attain. If the data in hand do not provide sufficient precision, the fact that the effect is identified in our model tells us that we need only collect more of the same data. We summarize this point in Reasoning Box 10.1.

Identification Problems and Remedies

Let's return now to our hand-crafted rocking chair example. It may be tempting to extend the reasoning we presented for identifying the probability of rolling a 3 for a die to our initial example concerning the rocking chairs. Just as confidence intervals "collapse" around p as the sample size increases, we may believe that confidence intervals will also "collapse" around β as we get more data on Sales and Price. While it is true that confidence intervals will generally shrink as the sample size increases, it is not necessarily true that they can or will close in on the right number in all cases. Sometimes, the data are such that, no matter how large a sample we take from the population, we can close in on the right number only by making a crucial assumption(s) about the data-generating process and/or by sampling from an expanded or alternative population.

297

In this section, we discuss two primary circumstances in which identification problems typically arise: when attempting to extrapolate and/or interpolate, and when there is variable co-movement in the population.

EXTRAPOLATION AND INTERPOLATION

LO 10.2 Explain extrapolation and interpolation and how each inherently suffers from an identification problem.

In our rocking chair example, note that given the structure of shipping charges, there are two ranges of prices that consumers ever experience. If the designer posts a price of \$200 for a chair, the end price will be either between \$210 and \$225, or between \$275 and \$300, depending on whether or not the customer lives in Wisconsin. A scatterplot for a data sample of Sales and Prices would look something like [Figure 10.2](#).

Suppose we wanted to learn how Sales change with a change in Price. Looking at [Figure 10.2](#), the data paint a pretty clear picture as to how these variables move together in the price range of \$210 to \$225 and in the price range of \$275 to \$300. However, we might want to know how Sales move

with Prices in other price ranges. Suppose the designer was considering an increase in the base price of \$200 to a price of \$235. In that case, he would first be interested in how Sales move with price in the price range of \$245 to \$260. Given that he does not observe these prices with his existing data, the only way he can form predictions about the relationship between Sales and Price in that range is to interpolate. **Interpolation** involves drawing conclusions where there are “gaps” in the data. Here, a **data gap** is any place where there are missing data for a variable over an interval of values, but data are not missing for at least some values on both ends of the interval. In our example, there is a data gap for prices between \$225 and \$275. Hence, to draw any conclusions about how Sales change with Price for prices in that range, we must interpolate.

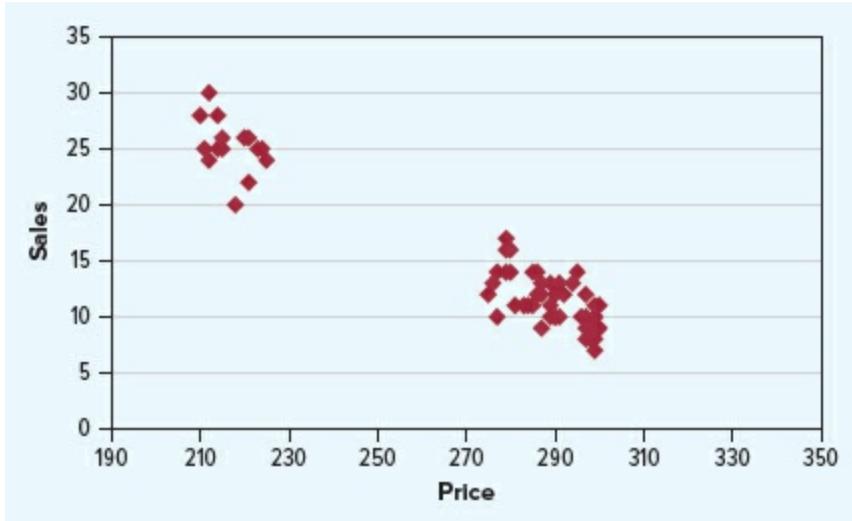
interpolation Drawing conclusions where there are “gaps” in the data.

data gap Any place where there are missing data for a variable over an interval of values, but data are not missing for at least some values on both ends of the interval.

A closely related issue to interpolation is that of extrapolation. We define **extrapolation** as drawing conclusions beyond the extent of the data. In our example, prices are never below \$210 and never above \$300. Consequently, to draw conclusions about how Sales relate to Price for prices below \$210 (below the lower extent of Price) or above \$300 (above the upper extent of Price) is to extrapolate.

extrapolation Drawing conclusions beyond the extent of the data.

FIGURE 10.2 Example of Possible Scatterplot of Sales/Price Data for Rocking Chairs



298

Identification Problems Identification problems are always something to consider when engaging in interpolation and/or extrapolation. The determining factor is whether the gap(s) in, or extent of, the data are due to random limitations in the sample or limitations in the population. If it is the former, there may be no identification problem—a larger sample of the same population may fill in the gap and/or increase the extent of the data. If it is the latter, then there is an identification problem that must be addressed.

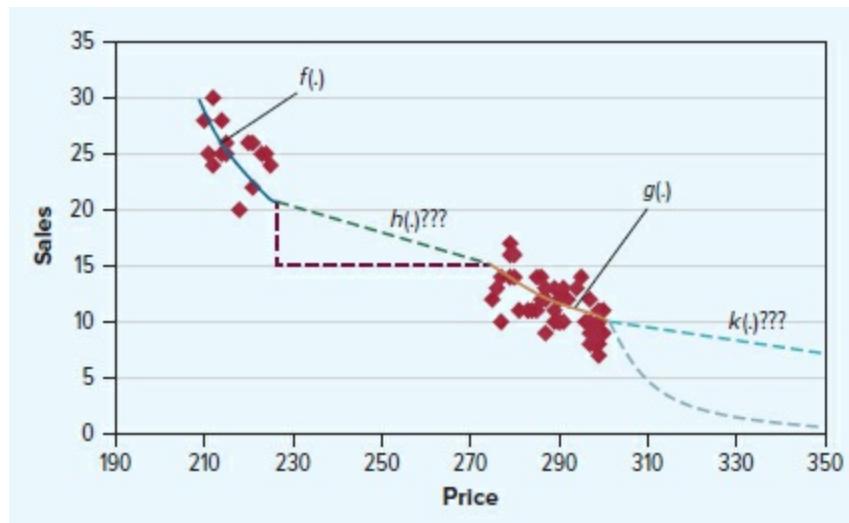
We can illustrate these points in our rocking chair example. To begin, suppose we want to make predictions about how Sales move with Price when Price varies between \$210 and \$225 and when Price varies between \$275 and \$300. For both of these price ranges, let's make minimal assumptions about the data-generating process by assuming $Sales_i = f(Price_i) + U_i$ for Price between \$210 and \$225, and $Sales_i = g(Price_i) + V_i$ for Price between \$275 and \$300. Here, we are not imposing any functional form for the relationship between Sales and Price, and we are allowing the functional form of this relationship to be different across the two price ranges. We do not detail how to conduct nonparametric estimation techniques in this book, which is what would be required if we wanted to estimate $f(\cdot)$ and $g(\cdot)$ without further assumptions. However, intuitively, the solutions we would get for $f(\cdot)$ and $g(\cdot)$

were we to utilize such methods are essentially what we get if, to the best of our ability, we draw freehand a curve that “best” fits the data—that is, generates residuals with mean zero and that are not correlated with Price. In Figure 10.3, we recreate Figure 10.2, but with attempts to draw $f(\cdot)$ and $g(\cdot)$ without any mathematical formulas.

Our attempt to draw $f(\cdot)$ and $g(\cdot)$ generates an important insight. If any two people were asked to draw $f(\cdot)$ and $g(\cdot)$ freehand within this scatterplot, the curves may not be exactly the same, but they will likely be similar. And if we added yet more data for each Price range, it will become even more likely that the curves separate people would draw look almost identical. This is the essence of $f(\cdot)$ and $g(\cdot)$ being identified within this population of data—as we get more data, their shape becomes less ambiguous. Even without assuming the shape of $f(\cdot)$ or $g(\cdot)$, we can estimate exactly what they look like given enough data.

Now, suppose we want to make predictions about how Sales move with Price when Price varies between \$225 and \$275 and when Price exceeds \$300. For both of these price ranges, let’s again make minimal assumptions about the data-generating process by assuming

FIGURE 10.3 “Freehand” Drawings of the Functional Relationship between Sales and Price



$\text{Sales}_i = h(\text{Price}_i) + W_i$ for Price between \$225 and \$275, and $\text{Sales}_i = k(\text{Price}_i) + A_i$ for Price above \$300. Suppose we asked different people to draw freehand curves estimating $h(\cdot)$ and $k(\cdot)$ that “best” fit the data, just as we did for the ranges (\$210,\$225) and (\$275,\$300). Both data ranges lack any data. Therefore, when drawing the curve estimating $h(\cdot)$, we are attempting to interpolate (fill in a data gap), and when drawing the curve estimating $k(\cdot)$, we are attempting to extrapolate (extend beyond the data’s range). In both cases, there is much more room for debate as to what curve best fits the data, since there are no data in either range for us to fit—only a starting and ending point. In [Figure 10.3](#), we present just a few of the many possible options (as dashed lines). Importantly, this problem is not due to our randomly having no data points in these ranges; even if we collected more and more data, the problem would not go away. No matter how much data we collect, it would still be the case that we have no data points in these ranges if we are sampling from the same population. Interpolation and extrapolation over ranges where no data are available in the population inherently suffer from an identification problem. Even just looking at our candidate curves in [Figure 10.3](#), there is no data-driven argument for one over the other, and collecting more of the same data will not change this fact.

To summarize, when interpolation or extrapolation is used to fill in gaps or limited extent of the data sample, but not the population, there is *not* an identification problem. In the example, suppose our lack of data on prices between \$225 and \$275 were due only to the chance draw of our data, and not because prices between \$225 and \$275 never happen in the population. Then we have reason to believe that collecting more data would allow us to converge on the determining function relating Sales to Price over this range, since we will eventually have observations with prices between \$225 and \$275. In contrast, when interpolation or extrapolation is used to fill gaps or limited extent of the population, there *is* an identification problem. No matter how much data is collected from the population, it will not help us to draw any conclusions about what is happening in the unobserved range(s).

Remedies Suppose we wish to engage in interpolation and/or extrapolation when there exists an identification problem. We know then for a general model of the data-generating process, where no assumptions are made about the determining function, we cannot “close in” on features of the determining function by simply sampling more data from the population. There are two key approaches toward solving this type of identification problem: changes in the population, and a functional form assumption.

Certainly the most straightforward way of remedying an identification problem for interpolation and/or extrapolation is to change the population from which you are sampling. To follow this remedy, you must find a way to alter the population so that there exist elements of the population where there were data gaps (for interpolation) or limits to the extent of the data (for extrapolation). For our rocking chair example, to solve the identification problem when trying to interpolate Sales for Prices between \$225 and \$275, we may request the designer to try different base prices. If he charges a base price of \$150, he will observe Sales for Prices between \$225 and \$250 for out-of-state zip codes (after adding the \$75 to \$100 for shipping). And if he also tries a base price of \$175, he will observe Sales for Prices between \$250 and \$275. The original population consisted of Location, Sales, and Prices for a fixed base price of \$200 and varying shipping costs. Here, we can change the population by altering the base price, allowing us to attain data that fill in the gap in the original dataset. Similarly, a change in the base price to something higher than \$200 (say, \$220) will allow us to address identification

LO 10.3 Distinguish between functional-form assumptions and enhanced data coverage as remedies for identification problems stemming from extrapolation and interpolation.

300

problems from extrapolation, by allowing us to observe Sales that correspond to prices above \$300.

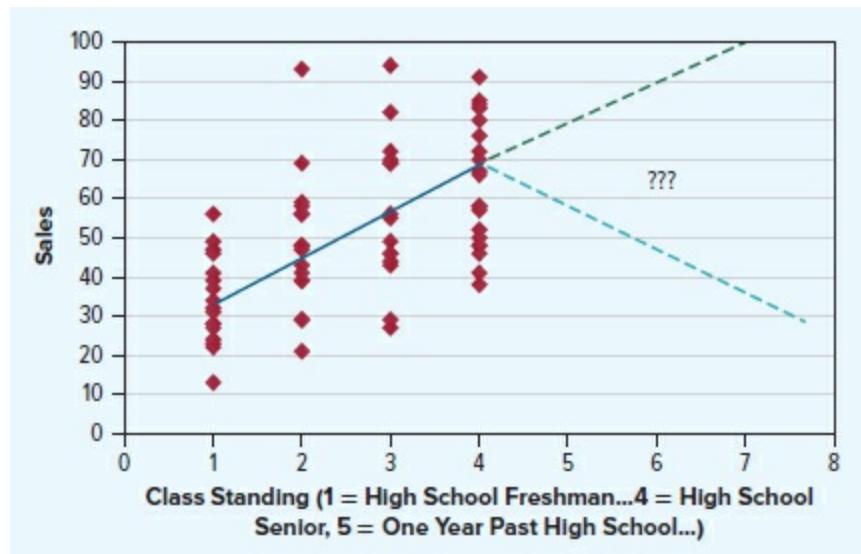
To further illustrate the applicability and potential importance of

changing the population to alleviate an identification problem, suppose a new pop singer is just starting her career and has been promoting her music by selling physical copies of her music at various high schools. She charges the same price to everyone, and has been taking note of differences in sales across high school classes. She finds that seniors buy the most often, freshman the least, and sophomores and juniors are in between. Just this fact alone tells her that her sales appear to be increasing by age of customers. Seeing this, she would like to extrapolate this relationship beyond just high school-aged kids. However, using only data from high schools, she has an identification problem. It may be tempting to believe sales will be even higher for an older crowd, but there is no support from the data for this belief, and continuing to sample from high schools cannot provide any. [Figure 10.4](#) illustrates possible ways we might extrapolate past age 18, but there are no data to sort through the options. A clear solution to this identification problem would be to try selling her music at colleges and collect data on her sales performance among this group. Hence, this simple expansion of the population will alleviate the identification problem she faced when trying to extrapolate beyond high school student ages.

The second approach toward remedying an identification problem for interpolation and/or extrapolation is to impose a functional form assumption. At the end of [Chapter 7](#), we noted that by assuming the form for the determining function, we impose a shape on the relationship between the outcome and treatment(s). We then discussed considerations in deciding which form to assume. Here, we note that an assumed functional form actually can do more than just impose a shape on variables' relationships—it can be used to fill in data gaps and/or extend beyond the limits of the data.

In our rocking chair example, as discussed previously, if we make no assumption about the form of the determining function in the regions where there are missing data, then virtually any curve we can draw (that is at least downward sloping) is a viable candidate as to how Sales relate to Price in those regions. However, standard practice in business (and all applications in this book) is to assume a functional form of the determining function

FIGURE 10.4 Music Sales by Class Standing



301

that applies for all relevant Price levels. As a simple example, we might assume a data-generating process with a linear functional form for the determining function:

$$\text{Sales}_i = \alpha + \beta \text{Price}_i + U_i$$

This assumption not only imposes the shape of the relationship between Sales and Price to be linear, but also dictates how to interpolate and/or extrapolate. In Figure 10.5, we recreate the data for this example (from Figures 10.2 and 10.3), along with the estimated regression equation if we assume a linear determining function. Here, we are *estimating* α and β using only data with Price in the ranges (\$210,\$225) and (\$275,\$300), but as can be seen in the figure, we are *applying* these estimated values across many other Price levels. We are using these values to interpolate between \$225 and \$275 and to extrapolate all the way to \$350.

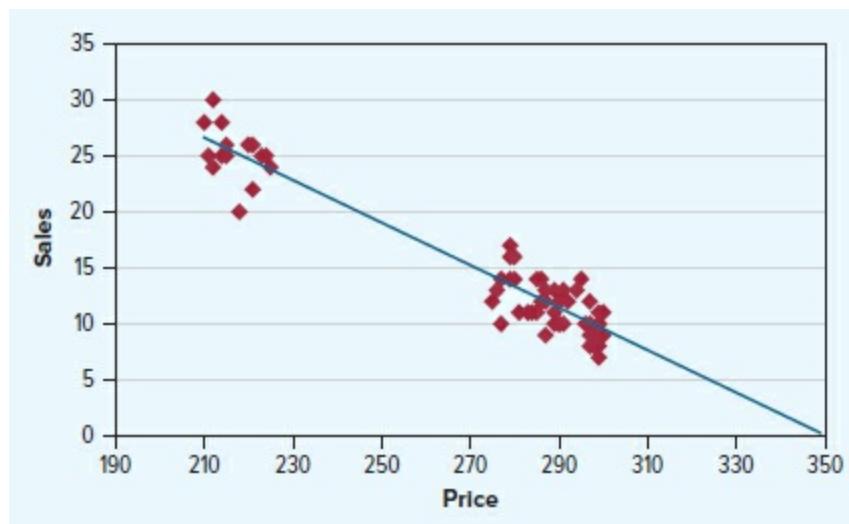
As can be seen in Figure 10.5, we can use the assumed form for the determining function to characterize the relationship between Sales and Price for price ranges we never observe. Hence, by assuming the form of the relationship between Sales and Price for *all* prices (at least between \$200 and

\$350 in the graph), we eliminate the identification problem for the prices we cannot observe. The data for prices in the ranges (\$210,\$225) and (\$275,\$300) can identify α and β , and through our functional form assumption, these values also apply for the price ranges of (\$225,\$275) and (\$300,\$350).

Unlike our first remedy (changing the population), particular care should be taken when using a functional form assumption as the sole basis for interpolation and/or extrapolation. While changing the population uses new or different data to fill in voids, and thus lets what we observe guide how we infer relationships between variables, a functional form assumption is just that—an assumption. We cannot use data to assess the applicability of this assumption for ranges where we have no data, and so we are left with theoretical arguments as to whether it is applicable for the voids we wish to fill.

For our rocking chair example, if we wish to interpolate the relationship between Sales and Price for prices between \$225 and \$275 by filling this gap with our assumed line, we must be prepared with theoretical arguments as to why this is reasonable. We might reference other demand studies indicating a linear relationship over a wide range of prices for

FIGURE 10.5 Regression Line for Rocking Chair Sales and Price Data



a similar product and argue (in theory) that this relationship should roughly extend to the product we are analyzing. Similarly, if we wish to extrapolate the relationship between Sales and Price for prices above \$300 using our line, we again must be able to make a theoretical argument. For extrapolation, we must also consider the extent to which we believe this functional form assumption applies. In our example, we are applying it up to prices of \$350, but do we want to apply it to even higher prices (e.g., \$400)? In general, the further the extent to which we wish to apply our functional form assumption, the stronger the theoretical support must be. This is because the range of values for which we are applying this assumption generally becomes more dissimilar to the values for which we actually have observations as we try to extrapolate further.

10.1

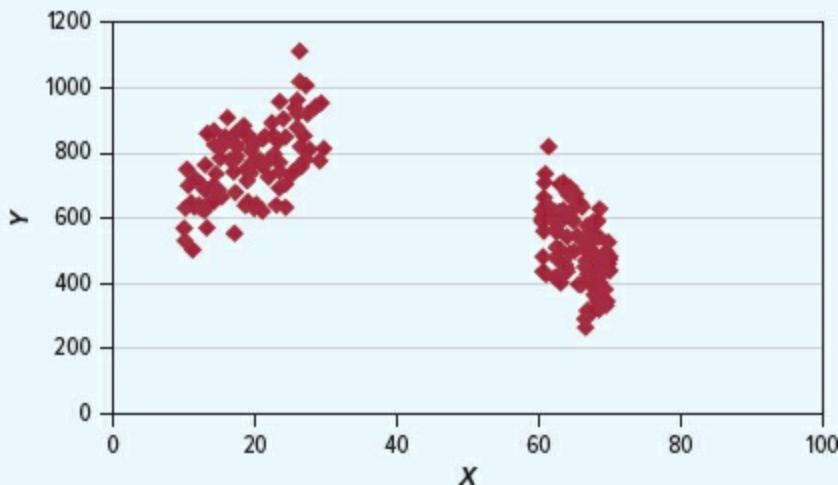
Demonstration Problem

You have collected data on an outcome (Y) and a treatment (X), where X takes on only values between 10 and 30 and between 60 and 70 in the population.

[Figure 10.6](#) plots the 200 observations that you have for Y and X .

FIGURE 10.6 Scatterplot of Data on Y and X

FIGURE 10.6 Scatterplot of Data on Y and X



- Is the relationship between Y and X identified for all values of X ?
- If you are unwilling to make a (potentially arbitrary) functional form assumption, how might you interpolate between 30 and 60 or extrapolate beyond 70?
- Suppose you are willing to make a functional form assumption so the assumed data-generating process is:

$$Y_i = \alpha + \beta_1 X_i + \beta_2 X_i^2 + U_i$$

Using regression analysis, you get the following estimated determining function:

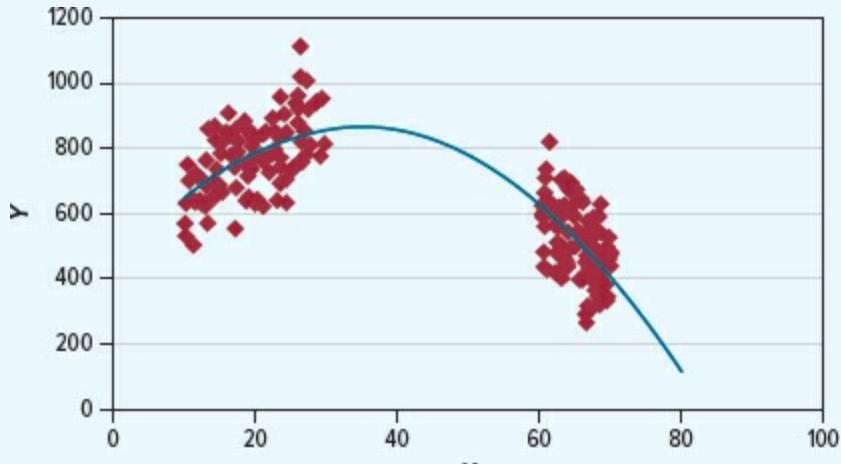
$$Y = 433.07 + 24.94X - 0.36X^2$$

303

- Draw this estimated determining function on [Figure 10.6](#) to show how it fills in voids in the data.
- For $X = 40$, according to your estimated determining function, what is the effect of an increase of X by 1 on Y ?

Answer:

- a. No. For X values less than 10, between 30 and 60, and over 70, the relationship between Y and X is not identified.
- b. Without a functional form assumption, you must find alternative data that contain values of X in these ranges.
- c. (i)



- (ii) The effect is $24.94 - 0.72 \times 40 = -3.86$. Note that this effect is in the interpolated range, and so we only have this number because of our functional form assumption (which we used to get it).

VARIABLE CO-MOVEMENT

The second circumstance in which identification problems typically arise is when there is variable co-movement in the population. We use the broader term “co-movement” rather than just correlation, since simple correlations alone do not encompass all the ways variables may move together in a population that result in identification problems. In this section, we discuss three types of variable co-movement: perfect multicollinearity, imperfect multicollinearity, and endogeneity. We begin by defining (or revisiting in the case of endogeneity) these terms. We then discuss if and how these types of variable co-movement lead to identification problems, and we follow this with discussion of possible remedies.

Consider the following data-generating process:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_K X_{Ki} + U_i$$

Now, suppose we want to use regression analysis to estimate $\alpha, \beta_1, \beta_2, \dots, \beta_K$. Here, we have assumed a functional form, so as long as there is some variation in X_1, X_2, \dots, X_K , there will not be identification problems stemming from voids in the data per se. However, we may still face an identification problem when there is co-movement among the X s and/or co-movement between one or more X and U .

304

COMMUNICATING DATA 10.1

PROJECTING TRENDS

While time series data are not the focal point of this book, some of the key insights and pitfalls pertaining to extrapolation often apply to time series data. It is often tempting to make projections based on current and past trends for an outcome of interest, but the simple fact is that predictions for any variable into the future require extrapolation and thus suffer from identification problems. For example, a firm may plot its profits over the past 10 quarters and arrive at [Figure 10.7](#).

FIGURE 10.7 Scatterplot and Regression Line for Firm Profits by Quarter

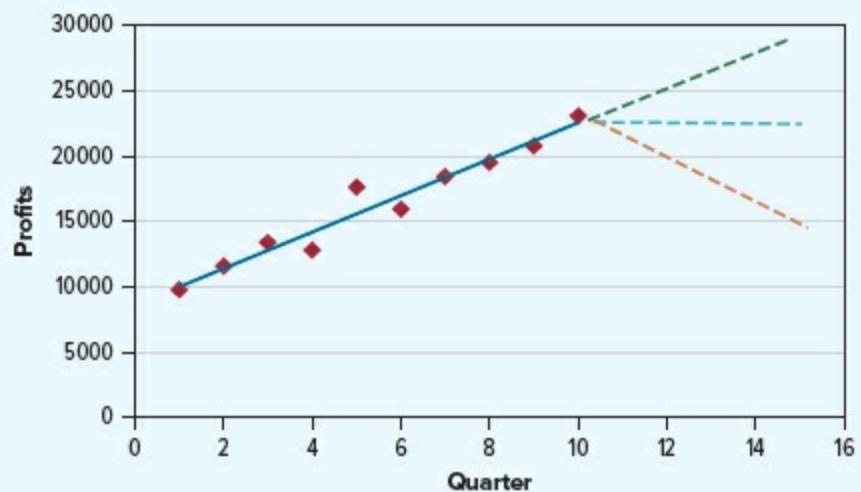


Figure 10.7 clearly shows a positive relationship between Profits and Time. However, even if we somehow believed Time per se improved Profits, any prediction we make about Profits beyond the present would completely rely on a functional form assumption. Each dashed line represents a possible future relationship between Profits and Time, and we have no choice but to wait for future data to sort between them with something other than theoretical arguments.

A particularly problematic type of co-movement among the independent variables (X s) is called perfect multicollinearity. **Perfect multicollinearity** is a condition in which two or more independent variables have an exact linear relationship. For example, if we can write $X_{1i} = c + dX_{2i}$, there is perfect multicollinearity in our model. More generally, perfect multicollinearity in our model is equivalent to being able to express $d_1X_{1i} + d_2X_{2i} + \dots + d_KX_{Ki} = c$ for all i in the population. Viewed a different way, perfect multicollinearity implies a special type of correlation among two or more independent variables. Perfect multicollinearity among a set of X s—say, X_1, \dots, X_J —implies that the semi-partial correlation between at least one pair of X s within this group, controlling for the other X s in the group, is equal to 1. So, if $J = 3$, perfect multicollinearity implies we have $\text{spCorr}(X_1, X_2 | X_3) = 1$, $\text{spCorr}(X_2, X_3 | X_1) = 1$, or $\text{spCorr}(X_1, X_3 | X_2) = 1$.

perfect multicollinearity A condition in which two or more independent variables have an exact linear relationship.

A second type of co-movement among the independent variables (X s) is called imperfect multicollinearity. **Imperfect multicollinearity** is a condition in which two or more

imperfect multicollinearity A condition in which two or more independent variables have *nearly* an exact linear relationship.

305

independent variables have *nearly* an exact linear relationship. When this condition exists for a data-generating process, we cannot express $d_1X_{1i} + d_2X_{2i} + \dots + d_KX_{Ki} = c$ for all i in the population. However, imperfect multicollinearity is equivalent to there being at least one semi-partial correlation that is “high”—nearly equal to 1. While there is no official cutoff for correlation being “high,” it is common to characterize a correlation above 0.8 as “high.” Thus, for the previous data-generating process, if we have, say, $\text{spCorr}(X_1, X_2(X_3 \dots X_K)) = 0.92$, we would say there is imperfect multicollinearity. As we will elaborate in the next section, in practice we seldom directly calculate semi-partial correlations to detect imperfect multicollinearity. Rather, we can make alternative calculations that are indicative of imperfect multicollinearity and are more intuitively linked to the problems imperfect multicollinearity can cause.

The third type of variable co-movement we consider in the context of identification problems involves co-movement between an independent variable(s) and the error term (unobservables) in a data-generating process. We consider the already-familiar concept of endogeneity, defined in [Chapter 7](#) as correlation between at least one X and U .

LO 10.4 Differentiate between endogeneity and types of multicollinearity as identification problems due to variable co-movement.

Identification Problems Now that we have defined (or redefined) perfect multicollinearity, imperfect multicollinearity, and endogeneity, we detail if and how each type of variable co-movement can lead to an identification problem. We begin with perfect multicollinearity. Put simply, perfect multicollinearity always leads to an identification problem in regression analysis. To see how, consider a slightly revised version of our rocking chair example. Suppose the shipping costs are completely determined by the distance from the designer's production facility and the customer's zip code, and in a linear way. Specifically, suppose shipping costs (in dollars) are: $\text{Ship}_i = 0.04 \times \text{Distance}_i$. Consequently, if the base price for a rocking chair is \$200, the final price will be $\text{Price}_i = 200 + 0.04 \times \text{Distance}_i$. Now, suppose we believe Sales of the rocking chairs depend not only on Price but also on Distance from the designer's location. We allow for dependence on distance since preferences for rocking chairs in general, or rocking chairs from our designer, may differ across customer locations, depending on how far they are from the designer's location.

Given the above insights, we assume the following data-generating process for rocking chair sales:

$$\text{Sales}_i = \alpha + \beta_1 \text{Price}_i + \beta_2 \text{Distance}_i + U_i$$

We'd like to use our data to estimate α , β_1 , and β_2 . Unfortunately, the population from which we are drawing suffers from perfect multicollinearity, creating an identification problem, particularly for β_1 , and β_2 . The presence of perfect multicollinearity is clear, since we can write one independent variable as a linear function of another for every element in the population: $\text{Price}_i = 200 + 0.04 \times \text{Distance}_i$. The identification problem comes from the fact that we cannot separately estimate β_1 and β_2 —the marginal effect of Price and Distance on Sales, respectively—because of the perfect linear relationship between Price and Distance. To see this, let's revisit the data-generating process, but plug in our expression for Price in terms of Distance. By doing so, we get:

$$\text{Sales}_i = \alpha + \beta_1(200 + 0.04 \times \text{Distance}_i) + \beta_2 \text{Distance}_i + U_i$$

$$\text{Sales}_i = (\alpha + \beta_1 200) + (0.04\beta_1 + \beta_2) \text{Distance}_i + U_i$$

306

This simple substitution highlights the identification problem. Since Price is a linear function of Distance, the data-generating process that appeared to depend on two variables really depends on only one, Distance. Through regression, we can see how Sales move with Distance, but this will inform us only about $0.04\beta_1 + \beta_2$, a combination of the marginal effects of Price and Distance, and will not allow us to estimate β_1 and β_2 separately. Viewed another way, if we regress Sales on Distance, we solve two sample moment equations (residuals have mean of zero, and residuals are uncorrelated with Distance); however, there are *three* parameters we'd like to estimate in the original determining function. This leaves us with two equations and three unknowns, which generally has an infinite number of solutions. Ultimately, our inability to separate β_1 and β_2 constitutes an identification problem, as no increase in this type of data will allow us to estimate these parameters separately.

In our modified rocking chair example, it was relatively easy to detect the presence of perfect multicollinearity since we knew the method for calculating Price linearly depended on Distance due to shipping costs. More generally, there are essentially three ways to detect perfect multicollinearity for a given data-generating process and population. The first, and most straightforward, is via a known linear relationship among two or more independent variables. This was the situation for our modified rocking chair example—we knew the base price was constant and shipping costs were a linear function of Distance, so Price linearly depended on Distance.

The second way of detecting perfect multicollinearity is by recognizing misuse of dummy variables. As noted in [Chapter 7](#), failure to choose a base group when using dummy variables to represent a categorical, ordinal, or interval variable generates perfect multicollinearity. In our rocking chair example, suppose we want to control for regional preferences. To do so, we

create a categorical variable, Region_i , that takes on exactly one of the values North, South, East, or West, depending on the location of observation i . To control for the categorical variable Region, we create dummy variables for each of its possible values, call them North_i , South_i , East_i , and West_i , where, for example, $\text{North}_i = 1$ if $\text{Region}_i = \text{North}$, and 0 otherwise. Now, suppose we tried to include all the dummy variables in our determining function, resulting in the following assumed data-generating process:

$$\text{Sales}_i = \alpha + \beta_1 \text{Price}_i + \beta_2 \text{North}_i + \beta_3 \text{South}_i + \beta_4 \text{East}_i + \beta_5 \text{West}_i + U_i$$

Here we have perfect multicollinearity among the variables North_i , South_i , East_i , and West_i . This is because we can write each dummy variable as a linear function of the others, e.g., $\text{North}_i = 1 - \text{South}_i - \text{East}_i - \text{West}_i$ (when each of South_i , East_i , and West_i is 0, $\text{North}_i = 1$, and when any of South_i , East_i , and West_i is 1, $\text{North}_i = 0$). Consequently, including all four dummy variables leads to an identification problem, as it is impossible, no matter how much data we collect, to separately estimate β_2 , β_3 , β_4 , and β_5 . For this reason, we must choose one of the dummy variables to represent the base group and exclude it from the determining function. Dropping one of the dummies breaks the perfect multicollinearity and allows for identification of all the remaining parameters.

The third way of detecting perfect multicollinearity is simply to let the data reveal it. If perfect multicollinearity exists and we attempt to execute multiple regression, the computer will be unable to produce estimates for all parameters no matter what statistical package we are using. In Excel, for example, perfect multicollinearity is readily apparent in regression

TABLE 10.3 Regression Results with Perfect Multicollinearity between Price and Distance

	COEFFICIENTS	STANDARD	t STAT	P-VALUE	
--	--------------	----------	--------	---------	--

		ERROR			
Intercept	119.4606254	2.046683426	58.36790581	1.5741E-58	
Price	0	0	65535	#NUM!	
Distance	-0.024964948	0.001224936	-20.38061682	#NUM!	

output. Suppose for our modified rocking chair example, where Price linearly depends on Distance, we ignore the fact that there is perfect multicollinearity. Thus, assuming $Sales_i = \alpha + \beta_1 Price_i + \beta_2 Distance_i + U_i$, we regress Sales on Price and Distance. [Table 10.3](#) contains a representative version of the output we would see from such a regression in Excel. The clear “red flag” that there is perfect multicollinearity in our model is the exact zero coefficient estimate for one of our parameters (Price in this case). Excel is unable to provide an estimate for all the parameters due to perfect multicollinearity, so it simply sets one of them to zero, effectively dropping its corresponding variable (Price in this case) from the determining function. In contrast, other statistical packages will drop one of the perfectly multicollinear variables before presenting results, but will effectively produce the same final outcome. We discuss dropping of multicollinear variables further in the next section.

Another potentially problematic type of co-movement among the independent variables (Xs) is imperfect multicollinearity. However, there is an important distinction between perfect multicollinearity and imperfect multicollinearity when it comes to identification. While perfect multicollinearity creates an identification problem, imperfect multicollinearity does not. As long as there is not an *exact* linear relationship among independent variables (and thus all semi-partial correlations are less than 1), it is possible to separately estimate the effects of each independent variable with enough data.

Although imperfect multicollinearity does not cause an identification problem, it can create challenges with inference. That is, imperfect multicollinearity can generate inflated p -values and confidence intervals, making it difficult to make any strong inductive arguments about population

parameters. Because there is not an identification problem, these challenges go away with enough data; however, it is not a given in practice that we can collect enough data to overcome them just by increased volume.

To illustrate imperfect multicollinearity and the challenges it presents, in our modified rocking chair example, suppose that rather than Price having a perfect linear relationship with Distance, Price has a near-perfect linear relationship with Distance. Suppose we have

$$\text{Price}_i = 200 + 0.04 \times \text{Distance}_i + V_i$$

where V_i contains other factors affecting shipping costs, such as local fuel prices, etc. Lastly, suppose the variance of V_i is relatively small (say, 2). This means that the Price for a given customer is mostly determined by Distance, and the value for V (which mainly ranges between -4 and 4) has a relatively small impact on Price. A customer at a Distance of 2,000 miles might have a value for V of 3 and so face a Price of $200 + 0.04 \times 2,000 + 3 = \283 . Another customer at a Distance of 400 miles might have a value for V of -2 and so face a Price of $200 + 0.04 \times 400 - 2 = \214 . Here, the difference in their Prices is \$69, and the vast majority of that difference (\$64) is due to their difference in Distance. In this example, Price and Distance have imperfect multicollinearity.

308

TABLE 10.4 Regression Results with Imperfect Multicollinearity between Price and Distance

	COEFFICIENTS	STANDARD ERROR	t STAT	P-VALUE	L
Intercept	442.3147	285.6049	1.548694	0.126311	-12
Price	-1.611633	1.428392	-1.128285	0.263347	-4.
Distance	0.03873	0.057373	0.675131	0.501987	-0.

Suppose we again want to determine the effect of Price and Distance on

Sales, and so we assume the following data-generating process:

$$\text{Sales}_i = \alpha + \beta_1 \text{Price}_i + \beta_2 \text{Distance}_i + U_i$$

There is not perfect multicollinearity, so unlike in [Table 10.3](#), we will be able to get estimates for all our parameters when regressing Sales on Price and Distance. In [Table 10.4](#), we present the results of a regression of Sales on Price and Distance for a hypothetical sample of 200 observations, where Price and Distance suffer from imperfect multicollinearity.

The first thing to notice in [Table 10.4](#) is that both Price and Distance have high p -values and confidence intervals that include 0. Hence, we would fail to reject that $\beta_1 = 0$, meaning we would fail to reject that Price has no effect on Sales. Similarly, we would fail to reject that $\beta_2 = 0$, meaning we would fail to reject that Distance has no effect on Sales. Of course, it could simply be the case that, despite what theory might tell us at least with regard to Price, we get these results because neither Price nor Distance affects Sales. However, it is also possible that the high p -values and wide confidence intervals for Price and Distance are due to their imperfect multicollinearity. When independent variables are imperfectly multicollinear, their close co-movement makes it difficult to distinguish the effect of one independent variable from another. In our example, the close co-movement between Price and Distance makes it difficult to determine whether movements in Sales are due to movements in one or the other. This difficulty manifests in greater uncertainty in our estimates; thus, our estimators have higher standard errors, leading to higher p -values and wider confidence intervals.

Fortunately, there are simple ways to check whether there is imperfect multicollinearity in a model, and thus the possibility that this condition is inflating our p -values and confidence intervals. As noted in the definition of imperfect multicollinearity, we could calculate semi-partial correlations among the independent variables and check whether they are close to 1. However, in practice, an alternative approach, known as the **variance inflation factor (VIF)**, is more commonly used. The VIF for an independent variable—

say, X_1 — is equal to $\frac{1}{1-R_{X_1}^2}$, where $R_{X_1}^2$ is the R -squared from regressing that independent variable (X_1) on all other independent variables (X_2, \dots, X_K) for a given determining function. Recall from [Chapter 6](#) that R -squared is the fraction of the total variation of X_1 that can be attributed to variation in the other X s (X_2, \dots, X_K).

variance inflation factor (VIF) For an independent variable—

say, X_1 — is equal to $\frac{1}{1-R_{X_1}^2}$, where $R_{X_1}^2$ is the R -squared from regressing that independent variable (X_1) on all other independent variables (X_2, \dots, X_K) for a given determining function.

We note here that, for any X , R_X^2 by itself is an appropriate measure of imperfect multicollinearity—as R_X^2 increases, there is a stronger linear relationship among the independent variables. In the extreme, as R_X^2 approaches 1, the relationship between the independent variables approaches a perfectly linear one and thus moves from imperfect multicollinearity to perfect multicollinearity. VIF is a simple function of R_X^2 , which also increases as the linear relationship among the independent variables strengthens.

309

However, in practice we use VIF as a measure for imperfect multicollinearity instead of R_X^2 because, while we have not explicitly presented the formulas for the standard errors of the regression estimators, it can be shown that each is a multiple of its VIF. Consequently, a doubling of an independent variable's VIF amounts to a doubling of the standard error of its coefficient estimator. A higher VIF for a given variable implies more noise (less certainty) in its coefficient estimator. Consequently, the VIF not only provides us information about the strength of the linear relationship among the independent variables, but also tells us how much uncertainty this co-movement in the X s is injecting into our estimators.

The last column of [Table 10.4](#) contains the VIF for both Price and

Distance in our hypothetical dataset. Notice that the VIF for Price (and Distance since there are only two Xs) is 1361.41. Simple algebra tells us that $R^2_{\text{Price}} = 0.999265$. Hence, there is a strong linear relationship between the independent variables in our regression, leading to imperfect multicollinearity. While there is no definitive cutoff for designating a VIF to be “high,” and thus indicative of imperfect multicollinearity, a popular choice of cutoff is 10. Thus, if a variable’s VIF is greater than 10, it suggests there is imperfect multicollinearity, which may lead to substantial uncertainty when estimating its coefficient. For the example in [Table 10.4](#), we have a rather extreme case of imperfect multicollinearity given our very high VIF.

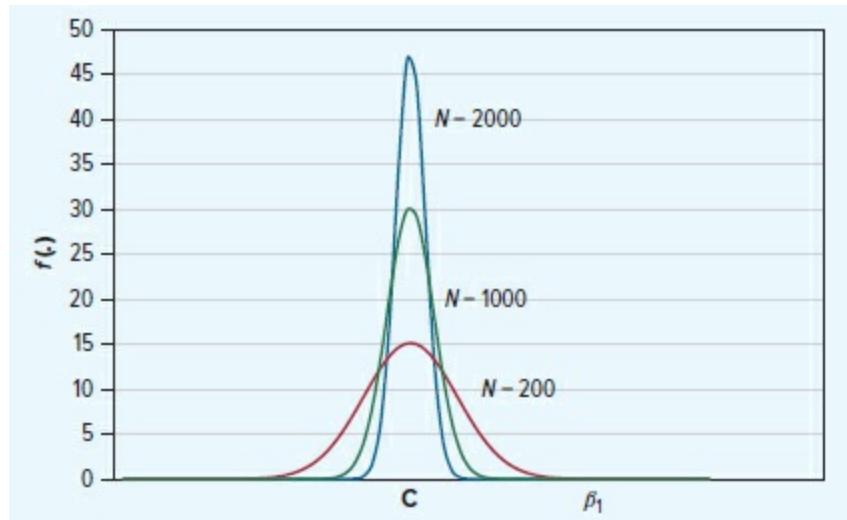
The third potentially problematic type of variable co-movement is the familiar condition of endogeneity. To this point, we’ve viewed endogeneity as problematic because it can lead to estimators that are not consistent. Here, we simply view this issue through a different lens. In fact, a model generating inconsistent estimators generally has an identification problem. Suppose we assume the data-generating process

$$Y_i = \alpha + \beta_1 X_{1i} + \dots + \beta_K X_{Ki} + U_i$$

and there is non-zero correlation between X_1 and U . We know that this correlation means $\hat{\beta}_1$ from a regression of Y on X_1, \dots, X_K need not be consistent, meaning its realized value need not get very close to β_1 as the sample size gets large. The inconsistency of $\hat{\beta}_1$ due to endogeneity amounts to **endogeneity as an identification problem**. The fact that $\hat{\beta}_1$ might “close in” on something other than β_1 as the sample gets large due to endogeneity means that increasing the sample does not necessarily give us a more precise estimate of β_1 . We illustrate this point in [Figure 10.8](#) where we have $\hat{\beta}_1$ approach a number $C \neq \beta_1$ as the sample

endogeneity as an identification problem Inconsistency of an estimator due to endogeneity

FIGURE 10.8 Example of Inconsistent Estimator



310

gets large. As can be seen in the figure, the confidence interval generated by $\hat{\beta}_1$ gets smaller as N gets larger, but it is providing a more precise estimate of the wrong number.

To summarize, note that both perfect multicollinearity and endogeneity create an identification problem, while imperfect multicollinearity does not. Imperfect multicollinearity creates challenges similar to an identification problem, since it makes it challenging to get precise parameter estimates. The difference is that imperfect multicollinearity makes getting precise estimates hard, while perfect multicollinearity and endogeneity make it impossible. Perfect multicollinearity and endogeneity also have an important distinction, since the former requires an exact linear relationship, while the latter requires only correlation. We summarize the discussion of this section in Reasoning Box 10.2.

LO 10.5 Articulate remedies for identification problems and inference challenges due to variable co-movement.

Remedies From Reasoning Box 10.2, we know that perfect multicollinearity and endogeneity both create identification problems. In this section, we

consider ways to overcome identification problems from each of these two sources.

We begin with perfect multicollinearity. Suppose we have assumed the following data-generating process:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots + \beta_K X_{Ki} + U_i$$

Suppose further that X_1 is the treatment, while X_2, \dots, X_K all play the role of controls. If our determining function suffers from perfect multicollinearity, the remedy depends on whether the exact linear relationship among the X s involves the treatment (X_1) or not.

Consider first the case where the linear relationship involves only the controls (a subset of X_2, \dots, X_K). As a simple example, suppose we have $X_{2i} = 5X_{3i}$ for all i . In this case, we are unable to distinguish the effect on the outcome (Y) of X_2 from X_3 . The solution to this problem is quite straightforward: simply drop X_2 or X_3 from the model. In general, the remedy for perfect multicollinearity involving only controls is to drop one of the variables comprising the exact linear relationship. To see why this is effective, let's substitute for X_2 in our assumed data-generating process to get:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2(5X_{3i}) + \beta_3 X_{3i} + \dots + \beta_K X_{Ki} + U_i$$

REASONING BOX 10.2

THE EFFECTS OF VARIABLE CO-MOVEMENT ON IDENTIFICATION

For the data-generating process $Y_i = \alpha + \beta_1 X_{1i} + \dots + \beta_K X_{Ki} + U_i$: If there exists an exact linear relationship between at least two of the independent variables (X s), defined as perfect multicollinearity, then there is an

identification problem.

In contrast, if there is no exact linear relationship among the X s, it is always possible to distinguish the effects of the independent variables on the outcome (Y) with any level of precision with sufficient data, even if some X s exhibit imperfect multicollinearity.

If there is correlation between any independent variable and the error term, defined as endogeneity, then there is an identification problem, no matter whether the correlation is via an exact linear relationship or not.

311

Combining terms, we have:

$$Y_i = \alpha + \beta_1 X_{1i} + (5\beta_2 + \beta_3) X_{3i} + \dots + \beta_K X_{Ki} + U_i$$

We see that, if we regress Y on X_1, X_3, \dots, X_K the coefficient estimate for X_3 will serve as an estimator for $5\beta_2 + \beta_3$. Thus, we can only estimate a (linear) combination of the effects of X_2 and X_3 on Y . However, the coefficient estimate for X_1 in this regression (where X_2 is dropped) still serves as an estimator for β_1 . Consequently, we can still get a consistent estimate for the effect of the treatment (β_1) after dropping one of the variables contributing to perfect multicollinearity. In sum, as long as our goal is to estimate the treatment effect and we have no particular interest in distinguishing the effects of controls, dropping one of the control variables contributing to perfect multicollinearity is an effective remedy.

Consider now our second case of perfect multicollinearity, in which the linear relationship involves the treatment (X_1 in our assumed data-generating process above). As another simple example, suppose we have $X_{1i} = 4X_{2i}$. The solution is not as simple as dropping one of the variables comprising the exact linear relationship. Assuming we are interested in the effect of the treatment, we cannot drop X_1 . We also cannot drop X_2 or else we create an

endogeneity problem—dropping X_2 would move it to the error term, and X_2 is clearly correlated with X_1 . The only viable remedy when the treatment contributes to a perfect multicollinearity problem is to change the population from which you are sampling. You must find a way to alter the population so that the treatment varies in ways that are not an

10.2

Demonstration Problem

Congratulations! You have just finished publishing your latest novel. To sell it, your publisher has entered an exclusive relationship with an online vendor. To learn a bit more about demand, the vendor has decided to choose a set of select locations to promote your book with a front-page ad and also will offer the book at a \$10 discount in those locations. Hence, in some locations the book is offered at a price of \$39.99 with no promotion; in other locations, the book is offered at a price of \$29.99 with a promotional ad. The vendor then collects data on the first-week sales across all locations. Assuming the data-generating process of

$$\text{Sales}_i = \alpha + \beta_1 \text{Price}_i + \beta_2 \text{Promotion}_i + U_i,$$

the vendor would like to estimate the effects of Price and Promotion on Sales. Note that Price takes on the values \$29.99 and \$39.99 in the data, and Promotion equals 1 if there was an ad and 0 otherwise.

- a. Using the data described above, what will happen if you regress Sales on Price and Promotion?
- b. If you drop Price from the regression, what will your coefficient estimate for Promotion actually be estimating?
- c. What can you do to separately identify the effects of Price and Promotion?

- a. For such a regression, you will be unable to get estimates for both β_1 and β_2 , since there is an exact linear relationship between these two variables. Specifically, we can write $\text{Price}_i = 39.99 - 10 \times \text{Promotion}_i$ for all i , meaning our data suffer from perfect multicollinearity. Consequently, your statistical software will either drop one of these two variables or produce a result as in [Table 10.3](#).
- b. Using the exact linear relationship between Price and Promotion, we can substitute for Price to get:

$$\text{Sales}_i = \alpha + \beta_1(39.99 - 10\text{Promotion}_i) + \beta_2\text{Promotion}_i + U_i = (\alpha + 39.99\beta_1) + (-10\beta_1 + \beta_2)\text{Promotion}_i + U_i$$

Consequently, regressing Sales on Promotion gives us an estimate of $-10\beta_1 + \beta_2$. In words, the coefficient estimate for Promotion would be an estimate of negative 10 times the effect of Price (the effect of a \$10 decline in Price) plus the effect of the Promotion.

- c. Dropping one of the variables will not allow us to separately identify the effect of Price and Promotion. Instead, we must try to acquire alternative data. For example, the vendor may try choosing some locations where it promotes the book without a price cut or vice versa. Alternatively, the vendor could vary the amount of the price cut for the locations that get a price cut and promotion.

exact linear function of other variables affecting the outcome. In our example, we need a population such that there are instances where $X_{1i} \neq 4X_{2i}$.

As we noted in the previous section, imperfect multicollinearity does not create an identification problem. Hence, if data are suffering from noisy estimates and VIF calculations suggest imperfect multicollinearity, the simple solution is to gather more data. However, depending on whether the imperfect multicollinearity involves just the controls (a subset of X_2, \dots, X_K

in our example) or the controls and the treatment (X_1 and a subset of X_2, \dots, X_K), gathering more data to address it may or may not be worthwhile. If the imperfect multicollinearity involves only the controls, and there is no interest in estimating the effects of the controls per se, then collecting more data will not necessarily be worthwhile. In this case, the estimated effect of the treatment is not being made noisier due to imperfect multicollinearity (since it is not part of the near-linear relationship), and so it is reasonable to stick with the data in hand, recognizing that imperfect multicollinearity is causing noisy estimates for the controls only. In contrast, if the imperfect multicollinearity involves the treatment, collecting more data likely will be worthwhile. More data will allow us more opportunities to observe the treatment move in ways different from the controls, and thus allow us to get a more precise estimate of the treatment's effect despite its imperfect multicollinearity with some (or all) of the controls.

We now turn to remedies for the other type of variable co-movement causing identification problems—endogeneity. The only viable remedy for endogeneity is to change the population from which you are sampling. And unlike the case of perfect multicollinearity, it does not matter whether the endogeneity involves the treatment or not. This is because correlation between any of the Xs (treatment or control) and the error (U) compromises

COMMUNICATING DATA 10.2

DISENTANGLING PROMOTION FROM FINANCING

A national chain of car dealerships is attempting to determine the effect on Profits of both television Promotion and the offer of 0 percent Financing for its vehicles. It collects cross-sectional data, covering one month, across all its dealerships on Profits, whether there was television Promotion of its vehicles and whether 0 percent Financing was available. You assume the data-generating process is

$$\text{Profits}_i = \alpha + \beta_1 \text{Promotion}_i + \beta_2 \text{Financing}_i + U_i$$

and the results of a regression of Profits on Promotion and Financing are shown in [Table 10.5](#):

TABLE 10.5 Regression Results for Profits Regressed on Promotion and Financing

	COEFFICIENTS	STANDARD ERROR	t STAT	P-VALUE
Intercept	116188.8344	1866.47432	62.25043288	3.38787E-8
Promotion	14802.88434	9964.246184	1.485600021	0.1405624
Financing	1569.335538	10003.01772	0.15688621	0.8756540

At first glance, it appears neither strategic variable (Promotion nor Financing) has a (statistically significant) impact on Profits. However, before arriving at this conclusion, you might ask whether these two variables are highly correlated—whether there is an imperfect multicollinearity issue with the data. Upon further examination, you find that the VIF for Promotion and Financing (which are the same, since there are only two independent variables) is 13.13. This is quite high (higher than 10), suggesting imperfect multicollinearity. It appears, then, that the car dealerships are very often engaging in Promotion at the same time they offer 0 percent Financing. This strong co-movement makes it difficult to disentangle each variable's individual effect on Profits.

One solution is to simply collect more data, which would allow for more instances in which Promotions and Financing offers don't occur at the same time. Another solution is to try to collect different data; the chain may try to randomly vary Promotions and Financing offers in a subsequent month across dealerships. This temporary policy change would greatly reduce the co-movement between Promotions and Financing and provide a good opportunity to disentangle their effects with relatively less data.

the consistency of all the estimators $(\hat{\alpha}, \hat{\beta}_1, \dots, \hat{\beta}_K)$. Consequently, we cannot count on the collection of more of the same data to help us “close in” on the corresponding parameters, no matter which X s are correlated with U .

The types of additional or alternative data we might collect are already familiar, as we’ve already discussed several ways of addressing endogeneity in prior chapters. Options include: collecting controls, finding a proxy variable(s), finding an instrument(s), and/or transforming cross-sectional data to become a panel. As can be seen from the discussion in this chapter, endogeneity essentially creates an identification problem, and all of these remedies for endogeneity serve to remedy this corresponding identification problem.

314

Identification Damage Control: Signing The Bias

LO 10.6 Solve for the direction of bias in cases of variable co-movement.

In some instances, we may recognize there is an identification problem that requires a change in the sampled population but be unable to acquire the necessary alternative data to address the problem. This leaves us in the difficult position of being forced to make use of what we know to be an inconsistent, or even inestimable, estimate of our treatment’s effect on the outcome. Unfortunately, if there is perfect multicollinearity involving the treatment and alternative data are not accessible, nothing can be done—there is simply no way to gain meaningful insight as to how the treatment, by itself, affects the outcome. However, if there is endogeneity and alternative data are not accessible, it is still possible to learn *something* about the treatment’s effect on the outcome under some circumstances.

Recall from [Chapter 7](#) that endogeneity is due to one of the following (potentially overlapping) causes: omitted variable(s), measurement error, and/or simultaneity. Recall also that the focus of our discussion of

endogeneity has been on the omitted variable form of endogeneity (though much still applies to measurement error and/or simultaneity). Continuing our focus on omitted variables, in this section we explain how, in some cases, it is possible to assess the direction of the inconsistency of our estimators in the presence of omitted variables.

To ground our discussion, consider a classic model of demand, where we assume unit sales depend on price in our data-generating process:

$$\text{Sales}_i = \alpha + \beta \text{Price}_i + U_i$$

Here, β represents the causal effect of a change in Price on the number of units sold (Sales). As we've highlighted in previous examples of demand estimation, we should be wary of simply regressing Sales on Price to get an estimate for β —the estimator for β ($\hat{\beta}$) is very likely to be inconsistent. The primary reason for concern with this simple regression is the high likelihood that there is at least one omitted variable that serves as a confounding factor (a variable in the error term that is correlated with Price).

If we are able to get only Sales and Price data, then we are stuck with an inconsistent estimator, which leaves us unable to draw meaningful conclusions about the effect of Price on Sales based solely on our regression estimates. However, suppose we know the source of the endogeneity is an omitted variable that serves as a confounding factor, and we have some knowledge about that omitted variable. In our simple demand example, let the product we are selling be rain boots, and we believe the omitted variable serving as a confounding factor is local rainfall. We believe local rainfall both influences the Sales of our rain boots, and relates to the Price we charge for the boots.

Now that we have a sense of what the confounding factor is in our model, we can try to go one step further by characterizing its relationship with both the outcome and the

treatment. In our example, we can first ask how local rainfall likely affects Sales for rain boots. More formally, we can ask what the sign of β_2

would be if we explicitly included local rainfall in our assumed data-generating process:

$$\text{Sales}_i = \alpha + \beta_1 \text{Price}_i + \beta_2 \text{LocalRain}_i + V_i$$

315

Since we do not have data on local rainfall, we cannot estimate β_2 using data; however, we can use basic theory to guide our assessment of the sign of β_2 . For our example, the theoretical argument is very straightforward: If there is more rain locally, there is likely more need for rain boots, so holding Price constant, more rain should lead to higher Sales of rain boots. Based on this reasoning, we would conclude that $\text{sign}(\beta_2) > 0$.

Next, we can ask how local rainfall relates to allocation of the treatment, Price. For this relationship, we are interested in how the treatment was allocated in our sample, not the population (we explain this further, momentarily). More formally, we can ask what is the sign of $\hat{\delta}$ for the estimated regression equation: $\text{Price}_i = \hat{\gamma} + \hat{\delta} \text{LocalRain}_i$. Again, without data on local rainfall, we cannot use data to determine the sign of $\hat{\delta}$. However, we again can make a straightforward theoretical argument: If there is more rain locally, there is likely more need for rain boots and thus a higher willingness-to-pay for rain boots; therefore, a firm can and will charge a higher Price. Based on this reasoning, we would conclude that $\text{sign}(\hat{\delta}) > 0$.

Before translating the above insights into a conclusion about β_1 , it is useful to link our example back to our original discussions of nonrandom treatments in [Chapter 4](#). Recall that inability to estimate an average treatment effect essentially stems from nonrandom treatment assignment, and in particular, treatment assignment that is correlated with other factors affecting the outcome. Hence, misestimating the treatment effect largely depends on another factor(s) affecting the outcome besides the treatment (for which we do not control) and this factor being correlated with the treatment assignment. Linking this reasoning to our example, β_2 represents how our “other factor” (local rainfall) affects the outcome, and $\hat{\delta}$ represents how this factor

correlates with the treatment assignment.

How does knowing the sign of β_2 and the sign of $\hat{\delta}$ help us to know more about β (the effect of Price on Sales)? This information tells us whether our estimate for β , $\hat{\beta}$, from regressing Sales on Price tends to overshoot or undershoot β . When Price is higher, there tends to be more local rainfall ($\text{sign}(\hat{\delta}) > 0$), and more local rainfall tends to generate more Sales of rain boots ($\text{sign}(\beta_2) > 0$); thus, the presence of more local rainfall with higher Prices generates more Sales than there otherwise would be with the Price increase. Therefore the presence of local rainfall as an omitted variable causes our regression estimate for the effect of Price ($\hat{\beta}$) to overshoot its true value (β).

Let's now generalize the intuition of our example. Suppose we have assumed the following data-generating process:

$$Y_i = \alpha + \beta_1 X_{1i} + \dots + \beta_K X_{Ki} + U_i$$

Let X_1 be the treatment and X_2, \dots, X_K be controls. Suppose also that there is an omitted variable, X_{K+1} , that affects Y (and so is part of U) and is correlated with X_1 . Consequently, an expanded version of the data-generating process can be written as:

$$Y_i = \alpha + \beta_1 X_{1i} + \dots + \beta_K X_{Ki} + \beta_{K+1} X_{K+1i} + V_i$$

Lastly, let $X_{K+1i} = \hat{\gamma} + \hat{\delta}_1 X_{1i} + \dots + \hat{\delta}_K X_{Ki}$ be the estimated regression equation we get if we were to regress X_{K+1} on X_1, \dots, X_K . Within this framework, define $\beta_{K+1} \times \hat{\delta}_1$ as the omitted variable bias. The **omitted variable bias** is the product of the effect of the omitted variable on the outcome (β_{K+1}) and the (semi-partial) correlation between the omitted variable and the treatment ($\hat{\delta}_1$). This product represents the difference between β_1 and the value that $\hat{\beta}_1$ (from regressing Y on X_1, \dots, X_K) consistently estimates. Put another way,

omitted variable bias The product of the effect of the omitted variable on the outcome (β_{K+1}) and the (semi-partial) correlation between the omitted variable and the treatment ($\widehat{\delta}_1$).

316

TABLE 10.6 Four Possibilities for the Sign of Omitted Variable Bias

		Sign of effect of omitted variable on the outcome	
		+	-
Sign of the (semi-partial) correlation between the omitted variable and treatment	+	+	-
	-	-	+

the omitted variable bias tells us how far “off” our estimate for the effect of X_1 tends to be from X_1 ’s true effect.

Since we do not observe the omitted variable, we cannot estimate either of the components of omitted variable bias. However, as was the case in our simple demand example, we often can use theory to guide us with regard to the *sign* of each component. And if we know the sign of each component, we know the sign of the omitted variable bias—that is, we know if $\widehat{\beta}_1$ tends to overshoot or undershoot β_1 . The basic relationship is:

$$\text{sign}(\beta_{K+1} \times \widehat{\delta}_1) = \text{sign}(\beta_{K+1}) \times \text{sign}(\widehat{\delta}_1)$$

In words, we have that the sign of the omitted variable bias is the product of the sign of the effect of the omitted variable on the outcome and the sign of the (semi-partial) correlation between the omitted variable and the treatment. We summarize the four possibilities for the sign of omitted variable bias in [Table 10.6](#), and the basic reasoning behind signing omitted variable bias in Reasoning Box 10.3.



REASONING BOX 10.3

SIGNING OMITTED VARIABLE BIAS

Suppose we have the following data-generating process:

$$Y_i = \alpha + \beta_1 X_{1i} + \dots + \beta_K X_{Ki} + U_i$$

where X_1 represents the treatment whose effect we are trying to estimate.

IF:

1. There is a single omitted variable contained in U (call it X_{K+1}) that is correlated with the treatment (X_1).
2. We can theoretically determine the sign of the (semi-partial) correlation between the omitted variable and the treatment—that is, we know $\text{sign}(\widehat{\delta}_1)$ where $\widehat{\delta}_1$ is the estimated coefficient on X_1 when we regress X_{K+1} on X_1, \dots, X_K .
3. We can theoretically determine the sign of the effect of the omitted variable on the outcome—that is, we know $\text{sign}(\beta_{K+1})$ where β_{K+1} is the effect of X_{K+1} on Y .

THEN:

When regressing Y on X_1, \dots, X_K , the bias for our estimated effect of X_1 is $\beta_{K+1} \times \widehat{\delta}_1$, and we know the sign of the bias is $\text{sign}(\beta_{K+1}) \times \text{sign}(\widehat{\delta}_1)$.

317

10.3

Demonstration Problem

Suppose you are consulting a large collective of farmers who sell corn across a large number of farmers' markets. These farmers have collected data on Profits and Prices for their corn for a given Saturday across many markets, and they

wish to learn the (average) effect of Price on Profits across these markets. To conduct your analysis, you assume:

$$\text{Profits}_i = \alpha + \beta \text{Price}_i + U_i$$

and regress Profits on Price. In doing so, you estimate the effect of a \$1 increase in Price on Profits is -\$500. However, you are suspicious of this estimate, as you fear Price is endogenous. To address this concern, you ask the farmers how they set Prices, and they indicate that weather is the key determinant: When weather is good, they Price higher, and when weather is bad, they Price lower.

- a. Based on this information, assess whether your estimate for β is likely too high or too low.
- b. Clearly explain *why* your estimate for β is biased in the direction it is when regressing Profits on Price.

Answer:

- a. In the problem description, it is clear that Weather is the omitted variable causing an endogeneity problem, where we might define Weather as a dichotomous variable, equaling 1 if weather is good and 0 if it is bad. We also know that Weather and Price are positively correlated. Lastly, it is reasonable to assume that, for a given Price, Profits will be higher when Weather is good and lower when Weather is bad. Combining this information, we can conclude we are in the upper left box of [Table 10.6](#), and so have a positive bias in our Price coefficient. Hence, the actual coefficient on Price is something smaller than -\$500 (i.e., Price's effect on Profits is more negative).
- b. Conceptually, the reason for the positive bias in our Price coefficient is as follows. When Price is higher, we know Weather is likely better (since Price and Weather are positively correlated). Further, we know that when Weather is better, Profits are higher. Putting these two facts together, an increase in Price generally corresponds with an increase in another variable that improves Profits. Since that other variable (Weather) is

unaccounted for (it is in the error term), our estimate for the effect of Price will include the actual effect of Price along with this additional positive effect on Profits from Weather.

COMMUNICATING DATA 10.3

A DISTORTED VIEW OF A DEGREE'S VALUE

Through clever measurement and accounting, suppose analysts at your firm were able to determine the marginal revenue product (MRP) of each employee—that is, the amount each employee adds to firm revenues each year. The analysts also have information on the institution from which each employee received his/her college degree. Using these data, they run a simple regression of MRP on Ivy, where Ivy is a dichotomous variable equaling 1 if the employee received a degree from an Ivy League school and 0 otherwise. The results of this regression show that an Ivy League degree increases MRP by \$23,000 on average.

318

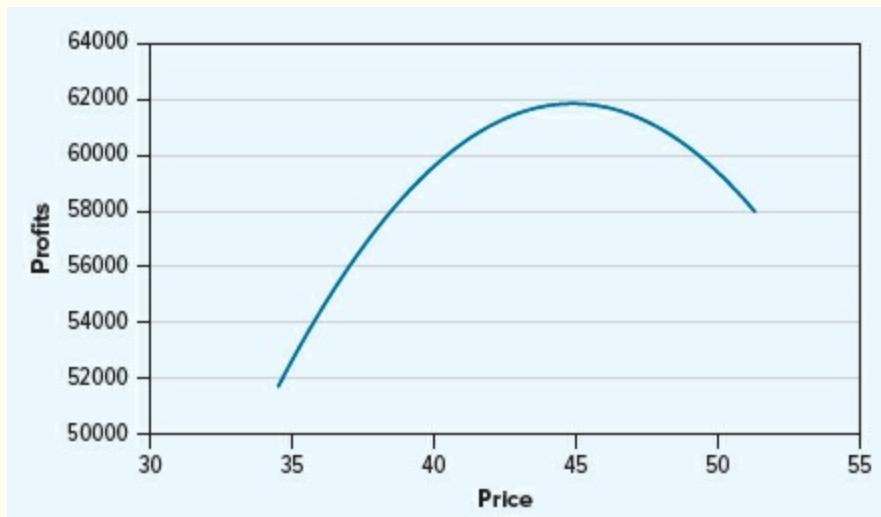
However, with a deep understanding of endogeneity problems, you note that this regression result is almost certainly an inaccurate estimate of the causal effect of an Ivy League degree on an employee's productivity. However, you can do more than just discredit this result; you can explain the likely direction in which it is "off." Specifically, employees with an Ivy League education likely have higher (unaccounted-for) raw intellect, and higher raw intellect likely results in greater MRP. Consequently, the estimated effect of an Ivy League education on MRP is likely overstated—it combines the true effect of an Ivy League education on MRP with the (positive) effect of raw intellect, which positively correlates with an Ivy League education.

RISING TO THE **data**CHALLENGE

Are Projected Profits over the Hill?

The data you used to get your regression results in [Table 10.1](#) had only prices between \$34.99 and \$38.99. Therefore, assessing the effect of raising price to \$49.99 requires extrapolation from the current range of data. Based on the regression results, it would appear that raising price from \$36.99 to \$49.99 would raise profits by \$19,548.62 ($= 1503.74 \times 13$). However, this prediction is not well grounded in the data since we never observed a price above \$38.99. Instead, it was our functional form assumption (that Profits are a linear function of Prices) that allowed us to predict how Profits would change for Prices above \$38.99. However, it is possible, and even likely, that Profits have a quadratic, or “hill-shaped,” relationship with Prices. For example, the relationship between Profits and Price may look like the shape in [Figure 10.9](#).

FIGURE 10.9 Hill-Shaped Relationship between Profits and Price



Suppose the true relationship between Profits and Price does in fact look as it does in [Figure 10.9](#). Since we can identify only the segment between \$34.99 and \$38.99

using our data, we likely would be unable to determine where is the “top of the hill”—where profits peak and then subsequently turn lower.

The prior discussion highlights the fact that any projection as to how profits will change with an increase in price to \$49.99 depends entirely on our functional form assumption. Thus, unless we are completely confident in this assumption, we should be cautious in making such predictions. As we’ve seen in this chapter, the other remedy besides the functional form assumption is to get data over a broader range. In this case, we would want to conduct our experiment using prices ranging at least as high as \$49.99.

SUMMARY

This chapter introduced the concept of identification and explained how to assess data through the lens of its ability to identify a parameter(s) of interest. We explained that an understanding of identification is crucial when trying to estimate the effect of a treatment on an outcome—there is no point in undertaking the analysis if the effect is not identified. We next highlighted two common situations in which identification problems often arise—attempts toward extrapolation/interpolation, and data with variable co-movement. For each of these circumstances, we discussed alternatives for remedying the corresponding identification problem. Lastly, we addressed the unfortunate case in which an identification problem exists in the form of endogeneity and cannot be remedied. We explained that under some circumstances, it is possible to ascertain the direction in which the treatment effect is misestimated.

KEY TERMS AND CONCEPTS

data gap

endogeneity as an identification problem

extrapolation

identified

imperfect multicollinearity

interpolation

omitted variable bias

perfect multicollinearity

variance inflation factor (VIF)

CONCEPTUAL QUESTIONS connect

1. Suppose you've assumed the following data-generating process:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + U_i$$

Describe what it means for β_1 to be identified in this model for a given population of data. (LO1)

2. Your firm has just developed a new product, and you would like to learn the average rating, R , this product would receive from the population of U.S. adults (on a scale of 1 to 7). To estimate R , you plan to collect random samples from the population of U.S. adults. (LO1)
 - a. Let X_i be a random variable equal to the rating given to the product by U.S. adult i . Also, define the sample mean for a sample of size N as: $\bar{X}_N = \frac{1}{N} \sum_{i=1}^N X_i$. Show that, for a given confidence level (e.g., 95%), the length of the confidence interval for R using a sample of N observations of X_i is decreasing with N .

320

-
- b. Is R identified if we sample from this population?
3. Suppose you have data containing values for Y and X that was sampled from a population where $X < 50$. When regressing Y on X assuming a linear functional form, you get a slope of 7.2. You are willing to make the assumptions that will allow for a causal interpretation, so you conclude that an increase in X from 30 to 70 would, on average, increase Y by 288 ($= 40 \times 7.2$). (LO3)
 - a. By making a prediction for a change in X beyond 50, what are you

attempting to do?

- b. Why should you be skeptical of predictions for Y for X values above 50?
 - c. On what are you relying to make a prediction for Y for an X value above 50?
 - d. Suggest an alternative approach toward estimating how Y would change when X increases from 30 to 70.
4. What is the difference between extrapolation and interpolation? (LO2)
 5. True or False: "Interpolation does not suffer from an identification problem since there exist data in the population both above and below the range over which we are trying to estimate an effect." (LO2)
 6. Why does imperfect multicollinearity *not* create an identification problem, but perfect multicollinearity does create an identification problem? (LO4)
 7. Suppose you've assumed the following data-generating process:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + U_i$$

You have collected a dataset, and you are also willing to assume that you have a random sample and that the errors are not correlated with the X s. Below are the regression results when regressing Y on the X s: (LO5)

	COEFFICIENTS	STANDARD ERROR	t STAT	P-VALUE	LOWER 95%
Intercept	415.59	212.44	1.96	0.051	-2.66
X_1	77.03	30.53	2.52	0.012	16.93
X_2	-67.57	60.78	-1.11	0.267	-187.23
X_3	-71.76	108.35	-0.66	0.508	-285.07
X_4	58.18	46.03	1.26	0.207	-32.45

- a. According to these results, is there evidence that any of the X s affects Y ?
 - b. Why might collection of more data be particularly useful for this analysis?
8. Explain the difference between perfect multicollinearity and endogeneity,

and how each leads to an identification problem. (LO4)

9. In an effort to determine the effect of internal budget reviews on profit performance, your firm has collected data across many branches on whether the branch went through a careful budget review last year (year 1) and its profit per unit produced this year (year 2). Your estimated regression equation is: (LO6)

$$\text{Profit-per-Unit}_{i2} = 27 + 1.8 \times \text{Review}_{i1}$$

- a. What are some possible omitted variables from this regression that could be generating an endogeneity problem, thus precluding us from drawing any causal interpretation from the regression?
- b. What is the likely sign of the bias in our estimated effect of a budget review?

321

QUANTITATIVE PROBLEMS connect

10. Your firm is attempting to learn the effectiveness of a newly developed television ad on its sales. To do this, it has randomly run the ad between 0 and 5 times during one week across a large number of television markets in the United States. It then recorded product sales for the following month for each market. To conduct the analysis, analysts at the firm have assumed the following data-generating process:

$$\text{Sales}_i = \alpha + \beta \text{Ads}_i + U_i$$

Regressing Sales on Ads yields $\hat{\beta}$. The firm would like to use this number to project the change in Sales when increasing weekly television ads to 20. (LO3)

- a. According to these results, what is the expected change in Sales when Ads increase from 5 to 20?
- b. Why should we be skeptical of our result from Part a?

- c. What can you do to find an estimate of the effect of increasing Ads from 5 to 20 that is more credible?
- 11.** Your firm has just launched a new product and has solicited a large number of potential customers to rate its effectiveness between 0 and 100. You are particularly interested in how an individual's rating of the new product depends on his/her age. To estimate this effect, you divide the potential customers into age groups: under 25, 26–40, 41–55, and over 55. You then assume the following data-generating process:

$$\text{Rating}_i = \alpha + \beta_1 \text{Age25}_i + \beta_2 \text{Age2640}_i + \beta_3 \text{Age4155}_i + \beta_4 \text{Age55}_i + U_i$$

Dataset available at www.mhhe.com/prince1e

Here, each Age variable is a dichotomous variable that equals 1 if individual i belongs to that age group and 0 otherwise. Use the data in the file *Chap10Prob1112.xlsx*. (LO4)

- a. Create the dichotomous variables: Age25, Age2640, Age4155, and Age55.
 - b. Why are you unable to estimate the effect of each age group as listed?
 - c. Estimate and interpret the effect of changing age groups on the Rating.
- 12.** You've decided to expand your analysis from Problem 11, and you would like to learn how potential customers' ratings depend on both age and income. To perform this analysis, you've decided to treat both Age and Income as continuous, rather than categorical, variables. Consequently, you assume the following data-generating process: (LO5)

$$\text{Rating}_i = \alpha + \beta_1 \text{Age}_i + \beta_2 \text{Income}_i + U_i$$

Dataset available at www.mhhe.com/prince1e

Use the data in the file *Chap10Prob1112.xlsx*.

- a. Regress Rating on Age and Income, and comment on the significance of each independent variable.
- b. Provide evidence that there is imperfect multicollinearity in your

- regression results, and discuss the consequences.
- c. How might you remedy the imperfect multicollinearity that exists in this dataset?
- 13.** Your firm is attempting to determine the effect of sensitivity training on employee behavior. To do so, it has collected data for each employee on the number of times he/she has been reprimanded for insensitive behavior in the past year (Reprimands_i) and whether that employee received sensitivity training the prior year (Training_i). When regressing Reprimands on Training , the estimated effect of training is an average reduction in Reprimands of 0.21. (*LO6*)
- a. Argue why Training is likely endogenous in this regression.
 - b. What is the likely sign of the bias in your estimated effect of Training on Reprimands ?

**Data
Analysis
Critiques,
Write-ups,
and Projects**

APPLICATIONS

LEARNING OBJECTIVES

After completing this chapter, you will be able to:

- LO2.1** Critique data-driven conclusions.
- LO2.2** Produce a careful write-up of data analysis and active predictions.
- LO2.3** Produce a well-reasoned written and slide-oriented presentation making an active prediction.

Introduction

The purpose of Applications is to provide the opportunity to “put it all together”

after working through the chapters of the book. Each chapter builds a certain set of skills, but the real world generally requires us to use many of those skills in conjunction. By presenting several applications, with different points of emphasis and scope, we can see how and when to apply: different lines of reasoning, regression models, estimation methods, etc.

It is important to note that the applications we present are simplified versions of real-world problems. The simplifications allow us to maintain a focus on the skills presented in this book. And the insights one can gain by confronting the issues we present in these stylized examples generally carry through to unconstrained problems with even greater complexity.

323

Critical Analysis of Data-Driven Conclusions

LO A-1 Critique data-driven conclusions.

In this section, we present three examples of real-world conclusions driven by data analysis. We follow each presentation with a series of questions that probe the underlying reasoning when moving from the data to the conclusion.

CASE 1: TENNIS ANALYTICS

Tennis is a well-known game involving one-on-one competition. As with any competition, competitors choose their strategies of play, with each seeking strategies that will maximize the likelihood of winning. For tennis, a competitor has many strategies from which to choose. She must consider the placement of her serves, whether to concentrate her ground strokes toward her competitor's forehand or backhand, whether to play at the net or the baseline, etc.

In the run-up to a match, analysts commonly report various statistics for each competitor that might be informative toward the outcome of the match. For example, we might be given statistics such as those in [Table A.1](#) for one of the players—let's call her Serena—before the match.

The entries in [Table A.1](#) provide two measures and associated winning percentages. The first measure concerns the first serve percentage. In tennis, players alternate who serves the ball into play from one game to the next. For each point in a game, the server has two opportunities to serve the ball in the court, and if she fails on both occasions, she loses the point. Given she has two chances at success, the server typically is more aggressive on her first serve attempt, by hitting the ball harder and/or hitting it in a location that is tough for her opponent to reach. In [Table A.1](#), we see that when Serena is successful on at least 75% of her first serve attempts in a match, she wins the match 93% of the time; otherwise, she wins the match 76% of the time.

The second measure in [Table A.1](#) concerns the number of times Serena approaches the net in a point. At any time during a point, one (or both) of the players may come in close to the net in order to press her advantage and/or end the point quickly. The effectiveness of such a move can depend on the skill sets of both players, the element of surprise, etc. In [Table A.1](#), we see that when Serena approaches the net on more than 30 points in a match, she wins the match 91% of the time; otherwise, she wins the match 78% of the time.

After presenting [Table A.1](#), television tennis commentators typically voice their opinion on optimal strategy for the player in question, in this case Serena. Using these figures,

TABLE A.1 Winning Percentages Associated with Performance Statistics

MEASURE	WINNING %
First serve percentage > 75%	93
First serve percentage \leq 75%	76
Number of times at net > 30	91
Number of times at net \leq 30	78

the likely suggestions are clear. Key strategies for Serena are to (1) keep her first serves in play and (2) approach the net frequently.

1. The commentators are implicitly making two active predictions.
 - a. What are the active predictions they are implying?
 - b. What are the two treatments and associated outcome for these predictions?
 - c. Construct a simple expression for the data-generating process for the outcome for each treatment, where the outcome is a linear function of the treatment plus some unobservables.
2. Regarding assumptions:
 - a. What assumptions might we make that, if true, the commentators' claims about these two strategies would be valid?
 - b. How might these assumptions be violated for a given match?
 - c. Would your answer to 2a change if we knew the statistics in [Table A.1](#) were specific to her upcoming opponent (i.e., these figures were generated using only past matches between Serena and her upcoming opponent)?
3. If the commentators used the figures in [Table A.1](#) as a guide for people betting on the match, rather than to suggest strategy to Serena, would you view them differently?
4. Can you describe a set of circumstances where the information in [Table A.1](#) would lead to the predictions of the commentators? If so, detail the corresponding reasoning.

CASE 2: SWITCHING INSURANCE

In various advertising campaigns, insurance agencies commonly cite average savings among those who switched to their services. For example, we see claims such as: “People who switched saved \$582 on average.” Suppose this

figure is for Grade-A Insurance, a national car insurance company, and was calculated using all 12,832 people who switched to Grade-A in the past six months. Grade-A's intent in providing this statistic concerning switchers is apparent. It wants viewers of the advertisement to expect to save \$582 by switching to Grade-A, and so find it worthwhile to take the time and effort to investigate purchasing a Grade-A policy.

1. Grade-A's claim clearly intends for viewers of the advertisement to believe the $ATE = \$582$. What is the treatment and what is the outcome in this context?
2. Explain why treatment assignment is unlikely to be random in this example.
3. With random treatment assignment, we know $ATE = ETT$ and Selection Bias = 0; hence $ATE = ETT + \text{Selection Bias}$. In this example with likely nonrandom treatment assignment:
 - a. Should we expect $ATE = ETT$? Why or why not?
 - b. Should we expect Selection Bias = 0? Why or why not?

325

-
4. Construct a simple expression for the data-generating process for the outcome, where the outcome is a linear function of the treatment plus some unobservables.
 - a. What must be the value of the intercept in this expression?
 - b. Based on the information given in the example, what will be the estimated value of the slope in this expression?
 - c. Explain why the estimated value of the slope is likely a biased estimate for the causal effect of the treatment (i.e., the ATE).
 - d. What is the likely sign of the bias (positive or negative)? Explain.

CASE 3: GROCERY STORE PRICE PROMOTIONS

Salmond's is a chain of grocery stores operating in several states within the

United States Over the past two years, its stores have routinely used temporary price promotions to help boost product sales, particularly for their breakfast cereals. For example, stores typically have a “regular” price for their breakfast cereals, but then reduce the price by, say, 20% on occasion. Salmond’s is interested in determining the effects of these price promotions on revenues, and to do so, it hires a consulting firm, EKA Consulting, to help. It provides EKA with weekly data from 50 stores in Ohio and 56 stores in Texas on breakfast cereal sales and prices over a two-year period. Using the data it was given, EKA generates the following summary statistics in [Table A.2](#) (Note that Promotion equals one if cereal was on promotion in the observed store during the observed week).

ECM adds to these summary statistics by running a simple regression of Revenues on Price for the full sample. In particular, EKA assumes the following data-generating process for Revenues in store i during week t :

$$\text{Revenues}_{it} = \alpha + \beta \text{Price}_{it} + U_i$$

The regression results are in [Table A.3](#).

TABLE A.2 Summary Statistics for Salmond’s Breakfast Cereals

SAMPLE	VARIABLE	MEAN	STD. DEV.	MIN	MAX
Full sample	Revenues	9,997.34	4,145.28	196.59	23,519.05
	Average Price	3.06	0.34	2.19	4.12
	Promotion	0.49	0.50	0	1
Ohio only	Revenues	6,945.63	2,363.85	196.59	14,849.77
	Average Price	3.29	0.28	2.56	4.12
	Promotion	0.49	0.50	0	1
Texas only	Revenues	12,935.26	3,278.41	1,775.84	23,519.05
	Average Price	2.84	0.24	2.19	3.55
	Promotion	0.50	0.50	0	1

TABLE A.3 Regression Results for Revenues on Price for Salmond’s Breakfast Cereals



	COEFFICIENTS	STANDARD ERROR	t STAT	P-VALUE	LOWER 95%	UPPER 95%
Intercept	36643.90	773.98	47.34	0.00	35125.33	38162.47
Price	-8705.08	251.31	-34.64	0.00	-9198.15	-8212.01

326

Based on the results in [Table A.2](#) and [A.3](#), EKA notes that Salmond's could notably boost its revenues in Ohio if it charged prices in Ohio that were similar to those it was already charging in Texas.

1. With respect to EKA's claim, what is the treatment and what is the outcome?
2. It is highly unlikely that the difference in prices charged in Ohio versus Texas is due to a random pricing strategy. Although price is continuous, and not dichotomous, explain why differences in pricing strategy across states likely result in:
 - a. Selection Bias $\neq 0$.
 - b. ETT \neq ATE.

(Hint: Simplify the analysis by thinking of “high” prices vs. “low” prices.)
3. A team member at EKA recognizes the problems highlighted in Problem 2, and suggests you limit the analyses to the state level. For example, you can regress Revenues on Price only for Ohio. Again, although price is continuous, and not dichotomous, explain why differences in pricing strategy across time (within a state) likely result in:
 - a. Selection Bias $\neq 0$.
 - b. ETT \neq ATE.

(Hint: Simplify the analysis by thinking of “high” prices vs. “low” prices.)
4. Regarding prediction:

- a. Make a valid passive prediction using the results in [Table A.3](#), including the necessary assumptions.
 - b. Suggest changes to the assumed data-generating process and corresponding regression analysis that might produce results suitable for an active prediction.
-

Written Explanations of Data Analysis and Active Predictions

LO A-2 Produce a careful write-up of data analysis and active predictions.

In this section, we present three examples of regression results for mock data that are intended to answer real-world questions. We follow each set of results with a series of questions that require careful communication of what the results mean and their implications for active prediction.

CASE 1: INSURANCE CLAIMS AND DEDUCTIBLES

Advanced Auto offers baseline automobile coverage across several states of the Midwestern United States. The company generally offers a single policy option in all regions where it operates, but allows local managers to adjust the deductible for the policy they sell. Note that the deductible for a given policy represents the amount the policyholder must pay out of pocket before the insurance company begins paying a claim. For example, if a policyholder files a claim for \$1,500 and holds a policy with a \$400 deductible, the policyholder must pay \$400 and the insurance company pays \$1,100.

Advanced Auto is interested in learning how the choice of deductible for the policy it offers in a region affects the rate of claims filed by policyholders. To accomplish this task,

TABLE A.4 Summary Statistics for Advanced Auto Variables

VARIABLE	MEAN	STD. DEV.	MIN	MAX
Claims per 100	10.864	9.408	0	34.61
Deductible	496.4	282.961	0	1000
% Local Pop. 25–65	60.342	8.925	45	75
% Local Pop. Married	62.99	10.509	45	80
Local Traffic Index	5.638	2.866	1	10
Local Wealth Index	5.522	2.949	1	10

the company collected data from 500 separate markets that it serves at a given point in time. [Table A.4](#) provides summary statistics for the variables Advanced Auto collected.

The variables in [Table A.4](#) are as follows. “Claims per 100” is the number of claims per 100 policyholders that were filed the previous year in that market. “Deductible” is the deductible chosen by the local manager for the policy offered in that market. “% Local Pop. 25–65” is the percentage of the local population aged 25 to 65 years old. “% Local Pop. Married” is the percentage of the local population that is married. “Local Traffic Index” is a rating between 1 and 10 of the level of traffic on the roads in the local market, where 10 is a very high level of traffic and 1 is a very low level of traffic. “Local Wealth Index” is a rating between 1 and 10 of the wealth level of residents in the local market, where 10 is a very high level of wealth and 1 is a very low level of wealth.

To learn the effect of the policy’s deductible on the claim rate, analysts at Advanced Auto assumed the following data-generating process for the claim rate in market i :

$$Claims_i = \alpha + \beta_1 Deduct_i + \beta_2 Age25_26_i + \beta_3 Married_i + \beta_4 Traffic_i + \beta_5 Wealth_i + U_i$$

The analysts then regressed the claim rate on the deductible and the other variables in the assumed determining function. The results of this regression are in [Table A.5](#).

1. Describe the results in [Table A.4](#).
 - a. Is variation in Deductible important? If so, why?

2. Explain the results for Deductible in [Table A.5](#).
 - a. What does the point estimate mean?
 - b. What does the *p*-value mean?
 - c. What do the upper and lower bounds for the 95% confidence interval mean?

TABLE A.5 Regression Results for Claim Rate Regressed on Deductible and Other Variables

	COEFFICIENTS	STANDARD ERROR	t STAT	P-VALUE	LOWER 95%	UPPER 95%
Intercept	29.1421	1.7911	16.2704	0.0000	25.6229	32.66
Deductible	-0.0323	0.0007	-44.0612	0.0000	-0.0337	-0.03
% Between 25–65	0.0028	0.0170	0.1631	0.8705	-0.0307	0.036
% Married	-0.0352	0.0145	-2.4286	0.0155	-0.0636	-0.00
Traffic Index	0.0340	0.0574	0.5927	0.5537	-0.0788	0.146
Wealth Index	-0.0710	0.0541	-1.3112	0.1904	-0.1774	0.035

328

3. What purpose does including variables besides Deductible in the regression serve?
4. Are the variables besides Deductible that are in the regression all controls, or do any play the role of a proxy variable? What is the difference?
5. Predict the effect of raising the policy deductible by \$200 on claim rates.
6. Detail the line of reasoning necessary to make your prediction in

Question 5.

7. Identify at least one argument why your estimates may not be suitable for making the active prediction you made in Question 5. Be sure to highlight where your line of reasoning breaks down if this opposing argument is correct.

CASE 2: WEARABLE FEATURES AND SALES

WearRight recently released multiple models of its new wearable, the WristWrap, across a wide range of urban markets. Each model of the WristWrap tracks physical activity and monitors heart rate, among many other tracking and alert features; the key distinction across models is the battery life. For normal activity, the battery life across the different models ranges from 1 to 7 days.

Following the advice of EKA consulting, WearRight charged varying prices among battery-life models across many different markets. [Table A.6](#) provides summary statistics for the WristWrap's pricing and battery life across all models and markets. Here, Unit Sales/1,000 is wearable unit sales per 1,000 in the observed market.

WearRight wants to understand how unit sales depend on price and battery life, as well as the trade-off between these two product features. To address these issues, it asks EKA consulting to run an analysis. EKA assumes the following data-generating process for the unit sales in market i :

$$\text{UnitSales}_i = \alpha + \beta_1 \text{Price}_i + \beta_2 \text{BatteryLife}_i + U_i$$

EKA analysts then regressed Unit Sales (per 1,000) on the Price and Battery Life; the results of this regression are in [Table A.7](#).

TABLE A.6 Summary Statistics for the WristWrap

VARIABLE	MEAN	STD. DEV.	MIN	MAX
Price	131.84	18.06	103.53	161.42
Battery Life	4	2.24	1	7

Unit Sales/1,000	40.26	3.97	29	52
------------------	-------	------	----	----

TABLE A.7 Regression Results for Unit Sales (per 1,000) on Price and Battery Life

	COEFFICIENTS	STANDARD ERROR	t STAT	P-VALUE	LOWER 95%	UPPER 95%
Intercept	75.60	11.56	6.54	5.18E-10	52.80	98.40
Price	-0.36	0.12	-3.11	0.002	-0.59	-0.13
Battery Life	3.01	0.93	3.23	0.001	1.18	4.85

329

-
1. Describe the results in [Table A.6](#).
 - a. Is variation in Price for each model across markets important? If so, why?
 - b. Suppose instead that WearRight simply charged \$10 more for each 1-day increase in battery life in every market. How would that affect the regression results in [Table A.7](#)?
 2. Explain the results for Battery Life in [Table A.7](#).
 - a. What does the point estimate mean?
 - b. What does the p-value mean?
 - c. What do the upper and lower bounds for the 95% confidence interval mean?
 3. Explain the results for Price in [Table A.7](#).
 - a. What does the point estimate mean?
 - b. What does the p-value mean?
 - c. What do the upper and lower bounds for the 95% confidence interval mean?
 4. Predict the effect of raising the Battery Life by 1 day on Unit Sales per

1,000.

5. Detail the line of reasoning necessary to make your prediction in Question 4.
6. On average, how much is an extra day of Battery Life worth? (*Hint:* Determine how much the Price could increase along with a 1-day increase in Battery Life without reducing Unit Sales per 1,000.)
7. Suppose WearRight is concerned that there are inherent differences in market performance for the WristWrap, and the price/battery life combinations offered across markets are determined at least in part based on those differences.
 - a. How would this data feature damage the line of reasoning you presented in Question 5?
 - b. How might you append the data and the assumed data-generating process to remedy this problem?

CASE 3: AD DURATION AND CLICKS

Terricorps released its latest line of footwear and has been advertising its products on YouTube. In doing so, it has created two video ads with similar content but varying in duration. The two ad durations are 15 seconds and one minute. Terricorps is interested in whether, and how, the duration of the ad it shows affects the likelihood that the viewer clicks on the ad, taking the viewer to their website. To answer this question, Terricorps has hired EKA consulting and provided their analysts with data on individuals who viewed their ads, including whether the individual clicked the ad (Click), the duration of the ad viewed (Duration), the time of day (TOD), and whether there was congestion on the network used to view the ad (Congestion). All variables except TOD are binary.

To generate summary statistics, EKA converted ad duration into two dummy variables, one for each length (15 seconds and one minute). It also converted time of day into dummy variables for each category (Morning, Afternoon, Evening, and Night). Summary statistics for all the variables are

in Table A.8.

330

TABLE A.8 Summary Statistics for Viewers of Terricorps Ad

VARIABLE	MEAN	STD. DEV.	MIN	MAX
Click	0.199	0.400	0	1
Duration15	0.447	0.497	0	1
Duration60	0.553	0.497	0	1
Morning	0.207	0.405	0	1
Afternoon	0.294	0.455	0	1
Evening	0.403	0.490	0	1
Night	0.097	0.296	0	1
Congestion	0.542	0.498	0	1

To conduct its analysis, EKA assumes the following data-generating process for the binary variable Click:

$$Click_i = \alpha + \beta Duration60_i + \delta_1 Morning_i + \delta_2 Evening_i + \delta_3 Night_i + U_i$$

In constructing this equation, EKA omitted the dummy variable for 15 seconds and the dummy variable for Afternoon, using these as base groups for ad duration and time of day, respectively.

Before running the corresponding regression, EKA analysts recognized a possible problem. They noted that, using web tracking technology (e.g., cookies), longer ads were being shown to individuals who were more prone to buy Terricorps footwear based on their web browsing behavior. However, the analysts also noted that the level of congestion, even after controlling for time of day, affected the length of ad shown. Consequently, they ran a two-stage least squares analysis, using Congestion as an instrument for Duration60. Table A.9 contains the first stage and second stage results.

TABLE A.9 2SLS Results for Click on Ad Duration and Time of Day (Congestion as Instrument)

			STANDARD	P-	LOWER
--	--	--	----------	----	-------

		COEFFICIENTS	ERROR	t STAT	VALUE	95%
First Stage						
	Intercept	75.60	11.56	6.54	5.18E-10	52.80
	Morning	-0.032	0.020	-1.58	0.115	-0.071
	Evening	-0.014	0.018	-0.75	0.456	-0.050
	Night	-0.036	0.025	-1.42	0.156	-0.085
	Congestion	-0.374	0.021	-18.06	0.000	-0.414
Second Stage						
	Intercept	0.089	0.027	3.25	0.001	0.035
	Morning	-0.019	0.015	-1.24	0.216	-0.049
	Evening	0.014	0.015	0.94	0.347	-0.015
	Night	-0.025	0.019	-1.27	0.205	-0.063
	Duration60	0.201	0.044	4.59	0.000	0.115

331

-
1. Explain the results for Duration60 in [Table A.9](#).
 - a. What does the point estimate mean?
 - b. What does the *p*-value mean?
 - c. What do the upper and lower bounds for the 95% confidence interval mean?
 2. Explain the point estimates for the different times of day (Morning, Evening, Night).
 3. Is Congestion a valid instrument for Duration60?
 4. Predict the effect of using the one-minute ad versus the 15-second ad on the likelihood of a click.
 5. Detail the line of reasoning necessary to make your prediction in Question 4.
 6. Suppose EKA ran this analysis without using an instrument. That is, they simply regressed Click on the times of day and Duration60. Given just the information provided, what is the expected sign of the bias for

Duration60 in this regression?

Projects: Combining Analysis with Reason-Based Communication

LO A-3 Produce a well-reasoned written and slide-oriented presentation making an active prediction.

In this section, we present three projects requiring data analysis to provide active predictions for business strategy. For each project, we provide some basic background information and data, and then ask the analyst to predict the effect of a strategy change. It is then up to the analyst to make the prediction and support it with clear reasoning and data analysis, producing a report in written and slide-show format. The intended audience for the report is a firm manager with limited training in data analysis.

For each project, note the following:

1. The data are designed to be as realistic as possible. However, some features (e.g., variation of demographic measures) may seem unlikely to fully mirror what one would find if real data were collected. Such discrepancies exist only to facilitate the analysis in the more controlled environment we have created.
2. The instructions for each project are rather minimal, and the accompanying questions are very open-ended. This design is deliberate, as it provides the opportunity to explore various analytical approaches and lines of reasoning with minimal guidance—a scenario more closely tied to real-world problems.

PROJECT 1: TABLET PRICE AND PROFITS

TabletCo is experiencing lagging profits of late, and looking to make a strategic shift. One of the shifts the company is considering is a change in its

pricing. TabletCo only sells its product online, but the firm has been engaging in regional pricing, where it pegs the price it charges to the location of the buyer. To aid in its pricing decision, the company has collected monthly data on its tablets for the eight separate regions in which it sells them over the past four years. The data are in the file *Project1.xlsx*. The variables in the data file are listed and described in Table A.10.

Dataset available at www.mhhe.com/prince1e

TABLE A.10 Description of Variables in *Project1 .xlsx*

332

VARIABLE	DESCRIPTION
Month	Month during which observation was recorded
Region	Geographic region where observation was recorded
Population (in thousands)	Number of people living in the Region during the Month
Avg. Education	Average education level of people living in the Region during the Month
Avg. Income (in thousands)	Average income level of people living in the Region during the Month
Avg. Age	Average age level of people living in the Region during the Month
Avg. Household Size	Average size of all households in the Region during the Month
Unemployment	Unemployment rate in the Region during the Month
Rainfall	Recorded rainfall in the Region during the Month
Price	Average price charged for TabletCo tablet in the Region during the Month
Profit per capita	Profit per capita for TabletCo tablets in the Region during the Month

In your report and presentation:

- Predict what will happen if TabletCo raises or lowers its price.
- Detail the data-generating process you assumed to arrive at your results. Explain how you chose features of the determining function, e.g., its functional form and the variables in it.

3. Explain the estimation method you used to arrive at your estimates for the data-generating process.
4. After applying your estimation method, explain:
 - a. What do the point estimates mean?
 - b. What do the p -values mean?
 - c. What do the upper and lower bounds for the 95% confidence intervals mean?
5. Detail the line of reasoning necessary to make your prediction in Question 1.
6. Identify at least one argument why your estimates may not be suitable for making the active prediction you made in Question 1. Be sure to highlight where your line of reasoning breaks down if this opposing argument is correct.

PROJECT 2: AUTO AD BUDGET AND REVENUES

Universal Motors (UM) is hoping to both assess the effectiveness of its television ad campaigns and plan for future ad budgeting. UM has been running local ads in 400 different regions. The firm has collected cross-sectional data across all 400 regions, containing regional advertising expenditure, revenue per capita, and several other variables. The data are in the file *Project2.xlsx*. The variables in the data file are listed and described in Table A.11.

Dataset available at www.mhhe.com/prince1e

TABLE A.11 Description of Variables in *Project2 .xlsx*

VARIABLE	DESCRIPTION
Region	Geographic region where observation was recorded
Population	Number of people living in the Region during the Month (in

	thousands) in the past year
Avg. Education	Average education level of people living in the Region over the past year
Avg. Income	Average income level of people living in the Region (in thousands) over the past year
Avg. Age	Average age level of people living in the Region over the past year
Coast	Binary variable indicating whether the Region is on the East or West Coast (equals one if Yes and 0 if No)
Avg. Household Size	Average household size in the Region over the past year
Avg. Unemployment	Average unemployment rate in the Region over the past year
Avg. Rainfall	Average monthly rainfall in the Region
Ad Price Index	Index for the price of advertising in the Region in the past year
Budget per Capita	Advertising budget per capita in the Region in the past year
Revenue per Capita	Revenue per capita in the Region in the past year

In your report and presentation:

1. Predict what will happen if UM raises or lowers its budget per capita.
2. Detail the data-generating process you assumed to arrive at your results. Explain how you chose features of the determining function, e.g., its functional form and the variables in it.
3. Explain the estimation method you used to arrive at your estimates for the data-generating process.
4. After applying your estimation method, explain:
 - a. What do the point estimates mean?
 - b. What do the p -values mean?
 - c. What do the upper and lower bounds for the 95% confidence intervals mean?
5. Detail the line of reasoning necessary to make your prediction in Question 1.

- Identify at least one argument why your estimates may not be suitable for making the active prediction you made in Question 1. Be sure to highlight where your line of reasoning breaks down if this opposing argument is correct.

PROJECT 3: MACHINE MAINTENANCE AND QUALITY

Lawner, Inc., produces affordable lawn mowers in facilities across the United States, but due to some highly publicized defect incidents, its reputation has been taking a hit. One approach Lawner is considering to help reduce defects and therefore hopefully improve its reputation is increased machine maintenance in its production facilities. Of course,

334

TABLE A.12 Description of Variables in *Project3 .xlsm*

VARIABLE	DESCRIPTION
Month	Month during which observation was recorded
Facility	Facility where observation was recorded
Output	Units of output produced in the Facility during the Month
Avg. Machine Age	Average age of machines in use in the Facility during the Month
Avg. Daily Production Hours	Average number of hours during which production occurred in the Facility during the Month
Avg. Workers per Machine	Average number of workers per machine in use in the Facility during the Month
Avg. Worker Education	Average worker education in the Facility during the Month
Avg. Worker Age	Average worker age in the Facility during the Month
Avg. Worker Salary	Average worker salary in the Facility during the Month
Rainfall	Recorded rainfall in the area surrounding the Facility during the Month
Machine Maintenance Rate	Proportion of machines in the Facility receiving maintenance during the Month
Faulty Units (per)	Number of faulty units (per 1,000) produced in the Facility

1,000) during the Month

maintenance is costly, so Lawner wants to quantify the impact of increased maintenance before committing the resources to do it. To make this assessment, Lanwer has collected monthly data on its 12 production facilities over the past eight years. The data are in the file *Project3.xlsm*. The variables in the data file are listed and described in [Table A.12](#).

Dataset available at www.mhhe.com/prince1e

In your report and presentation:

1. Predict what will happen if Lawner raises or lowers its machine maintenance rate.
2. Detail the data-generating process you assumed to arrive at your results. Explain how you chose features of the determining function, e.g., its functional form and the variables in it.
3. Explain the estimation method you used to arrive at your estimates for the data-generating process.
4. After applying your estimation method, explain:
 - a. What do the point estimates mean?
 - b. What do the *p*-values mean?
 - c. What do the upper and lower bounds for the 95% confidence intervals mean?
5. Detail the line of reasoning necessary to make your prediction in Question 1.
6. Identify at least one argument why your estimates may not be suitable for making the active prediction you made in Question 1. Be sure to highlight where your line of reasoning breaks down if this opposing argument is correct.

GLOSSARY

A

active prediction The use of predictive analytics to make predictions based on actual and/or hypothetical data for which one or more variables experience an exogenous alteration.

association analysis Attempts to discover dependencies, generally in the form of conditional probabilities, between two or more variables in the data.

average treatment effect (ATE) The average difference in the treated and untreated outcome across all subjects in a population.

B

base group The excluded dummy variable among a set of dummy variables representing a categorical, ordinal, or interval variable.

business analytics The use of data analysis to aid in business decision making.

business strategy A plan of action designed by a business practitioner to achieve a business objective.

C

categorical variable Indicates membership to one of a set of two or more mutually exclusive categories that do not have an obvious ordering.

causal inference The process of establishing (and often measuring) a causal relationship between a variable(s) representing a cause and a variable(s) representing an effect, where a change in the cause variable

results in a change in the effect variable.

cluster analysis Groups observations according to some measure of similarity.

confidence interval A range of values such that there is a specified probability that they contain a population parameter.

confounding factor The component(s) of the error, U_i , that are correlated with a treatment(s), X .

consistent estimator An estimator whose realized value gets close to its corresponding population parameter as the sample size gets large.

constructing a representative sample The four steps that are to be followed in building a representative sample.

continuous random variable A variable that takes on an (uncountable) infinite number of values.

control variable Any variable included in a regression equation whose purpose is to alleviate an endogeneity problem.

cross-sectional data Data that provides a snapshot of information at one fixed point in time.

D

dashboard A graphical presentation of the current standing and historical trends for variables of interest, typically KPIs.

data A collection of information.

database Organized collection of data that firms use for analysis.

data gap Any place where there are missing data for a variable over an interval of values, but data are not missing for at least some values on both ends of the interval.

data-generating process (DGP) The underlying mechanism that produces the pieces of information contained in a dataset.

data mining Pattern discovery, typically in large datasets.

data sample A subset of a population that is collected and observed.

data sanity check for a regression A comparison between the estimated coefficient for an independent variable in a regression and the value for that coefficient as predicted by theory.

deductive reasoning Reasoning that goes from the general to the specific; also known as *top-down logic*.

degree of support (also called inductive probability) The degree of confidence in the conclusion resulting from the stated observation(s) for an inductive argument.

descriptive statistics Quantitative measures meant to summarize and interpret properties of a dataset.

determining function The part of the outcome that we can explicitly determine, $f_i(X_{1i}, X_{2i}, \dots, X_{Ki})$.

deterministic variable Variable whose value can be predicted with certainty.

dichotomous (or binary) dependent variable A limited dependent variable that can take on just two values, typically recorded as 0 and 1.

dichotomous treatment Two treatment statuses—treated and untreated.

difference-in-differences (diff-in-diff) The difference in the temporal change for the outcome between the treated and untreated group.

direct causal relationship A change in the causal variable, X , directly causes a change in the effect variable, Y .

direct proof Proof that begins with assumptions, explains methods of proof, and states the conclusion(s).

discrete random variable A variable that can take on only a countable number of values.

dummy variable A dichotomous variable (one that takes on values 0 or 1) that is used to indicate the presence or absence of a given characteristic.

dummy variable estimation Uses regression analysis to estimate *all* of the parameters in the fixed effects data-generating process.

E

effect of the treatment on the treated (ETT) Treatment effect for the group given the treatment.

elasticity The percentage change in one variable with a percentage change in another.

empirically testable conclusion A conclusion whose validity can be meaningfully tested using observable data.

endogeneity problem Correlation exists between the errors and at least one treatment.

endogeneity as an identification problem Inconsistency of an estimator due to endogeneity.

error term Represents “unobserved” factors that determine the outcome.

estimator A calculation using sample data that is used to provide information about a population parameter.

exogeneity as an instrumental variable A variable that has no effect on the outcome variable beyond the combined effects of all the variables in the determining function (X_1, \dots, X_K).

exogenously altered A variable in a dataset that changes due to factors outside the data-generating process that are independent of all other variables within the data-generating process.

expected value (or population mean) The summation of each possible realization of X_i multiplied by the probability of that realization.

experiment A test within a controlled environment designed to examine the validity of a hypothesis.

experimental data Data that result from an experiment.

extrapolation Drawing conclusions beyond the extent of the data.

F

fixed effects The controls for cross-sectional groups.

fixed effects model A data-generating process for panel data that

includes controls for cross- sectional groups.

H

homoscedasticity Variance constant across all values of X .

hypothesis A proposed idea based on limited evidence that leads to further investigation.

hypothesis test The process of using sample data to assess the credibility of a hypothesis about a population.

I

identified Can be estimated with any level of precision given a large enough sample from the population.

imperfect multicollinearity A condition in which two or more independent variables have *nearly* an exact linear relationship.

independent (random variable) The distribution of one random variable does not depend on the realization of another.

independent and identically distributed (i.i.d.) The distribution of one random variable does not depend on the realization of another and each has identical distribution.

indirect causal relationship A change in X causes a change in Y , but only through its impact on a third variable.

inductive reasoning Reasoning that goes from the specific to the general; also known as *bottom-up logic*.

instrumental variable In the context of regression analysis, a variable that allows us to isolate the causal effect of a treatment on an outcome due to its correlation with the treatment and lack of correlation with the outcome.

interpolation Drawing conclusions where there are “gaps” in the data.

interval variable Indicates membership to one of a set of two or more mutually exclusive categories that have an obvious ordering, and the difference in values is meaningful.

irrelevant variable Variables that do not affect the outcome.

K

key performance indicators (KPIs) Variables that are used to help measure firm performance.

L

lag information Information about past outcomes.

latent variable A variable that cannot be observed, but information about it can be inferred from other observed variables.

lead information Information that provides insights about the future.

least absolute deviations (LAD) Use the sum of the absolute value of the residuals as the objective function and solve for the slope and intercept that minimize it.

limit-violating prediction A predicted value for a limited dependent variable that does not fall within that variable's limits.

limited dependent model A dependent variable whose range of possible values has consequential constraints.

linear probability model Regression analysis applied to a dichotomous dependent variable.

337

linear regression The process of fitting a function that is linear in its parameters to a given dataset.

logic A description of the rules and/or steps behind the reasoning process.

logit model A latent variable formulation for a dichotomous dependent variable that assumes a Logistic(0,1) distribution for the unobservables.

M

marginal effect The rate of change in the probability of a dichotomous dependent variable equaling 1 with a one-unit increase in an independent variable (holding all other independent variables constant).

maximum likelihood estimation (MLE) Population-level parameters are

estimated using values that make the observed outcomes as likely as possible for a given model.

measurement error When one or more of the variables in the determining function (typically at least one of the treatments) is measured with error.

multi-level treatment When a treatment can be administered in more than one quantity.

multiple regression Solving for a function that best describes the data that implies the use of OLS (or equivalently, the sample moment equations).

N

nonexperimental data Data that were not produced using an experiment.

normal random variable A specific type of continuous random variable with a bell-shaped pdf.

null hypothesis The hypothesis to be tested using a data sample.

O

objective degree of support A degree of support that has a statistical foundation, making it more credible as compared to a subjective degree of support.

objective function A function ultimately wished to be maximized or minimized.

omitted variable Any variable contained in the error term of a data-generating process, due to lack of data or simply a decision not to include it.

omitted variable bias The product of the effect of the omitted variable on the outcome (β_{K+1}) and the (semi-partial) correlation between the omitted variable and the treatment ($\hat{\delta}_1$).

ordinal variable Indicates membership to one of a set of two or more mutually exclusive categories that do have an obvious ordering, but the difference in values is not meaningful.

ordinary least squares (OLS) The process of solving for the slope and intercept that minimizes the sum of the squared residuals.

outlier detection Small subsets of observations, if they exist, that contain information far different from the vast majority of the observations in the dataset.

P

p-value The probability of attaining a test statistic at least as extreme as the one that was observed.

panel data: The same cross-sectional units over multiple points in time.

partial correlation The partial correlation between X and Y is a measure of the relationship between these two variables, holding at least one other variable fixed.

passive prediction The use of predictive analytics to make predictions based on actual and/or hypothetical data for which no variables are exogenously altered.

pattern Any distinctive relationship between observations within the dataset.

pattern discovery The process of identifying distinctive relationships between observations in a dataset.

perfect multicollinearity A condition in which two or more independent variables have an exact linear relationship.

pivot table A data summarization tool that allows for different views of a given dataset.

pooled cross-sectional data The result of two or more unrelated cross-sectional datasets being combined into one dataset.

population The entire set of potential observations about which we want to learn.

population mean The summation of each possible realization of X_i , multiplied by the probability of that realization.

population parameter A numerical expression that summarizes some feature of the population.

population standard deviation (σ) The mean of the sample standard

deviation.

predictive analytics The use of data analysis designed to form predictions about future, or unknown, events or outcomes.

probability density function (pdf) A function used to calculate probabilities of individual outcomes for a continuous random variable.

probability function A function used to calculate probabilities of individual outcomes for a discrete random variable.

338

probit model A latent variable formulation for a dichotomous dependent variable that assumes a standard normal distribution for the unobservables.

proxy variable A variable used in a regression equation in order to proxy for a confounding factor, in an attempt to alleviate the endogeneity problem caused by that confounding factor.

Q

query Any request for information from a database.

R

R-squared The fraction of the total variation in Y that can be attributed to variation in the X s.

random sample A sample where every member of the population has an equal chance of being selected.

random variable Variable that can take on multiple values, with any given realization of the variable being due to chance (or randomness).

reasoning The process of forming conclusions, judgments, or inferences from facts or premises.

regression analysis The process of using a function to describe the relationship among variables.

regression line for a dichotomous treatment For a dichotomous treatment, the line describing the relationship between the treatment and outcome by using the means for each treatment status.

relevant as an instrumental variable A variable that is correlated with X_1 after controlling for X_2, \dots, X_K .

report Any structured presentation of the information in a dataset.

representative sample A sample whose distribution approximately matches that of the population for a subset of observed, independent variables.

residual The difference between the observed outcome and the corresponding point on the regression line for a given observation.

robustness The persistent accuracy of a conclusion despite variation in the associated assumption(s) within the context of a deductive argument.

S

sample mean A common measure of the center of a sample.

sample moment The mean of a function of a random variable(s) for a given sample.

sample of size N A collection of N realizations of X_i , i.e., $\{X_1, X_2 \dots X_N\}$.

sample standard deviation The square root of the sample variance. For a sample of size N for random variable X_i , is

$$S^2 = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})^2}.$$

sample statistic Single measures of some feature of a data sample.

sample variance Common measure of the spread of a sample. For a sample of size N for random variable X_i , is $S^2 = \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})^2$.

scorecard Any structured assessment of variables of interest, typically KPIs, against a given benchmark.

scientific method The process designed to generate knowledge through the collection and analysis of experimental data.

selected sample A sample that is nonrandom.

selection bias The act of drawing conclusions about a population using a

selected data sample, without accounting for the means of selection.

semi-partial correlation The semi-partial correlation between X and Y is a measure of the relationship between these two variables, holding at least one other variable fixed for only X or Y .

simple regression The process that produces the simple regression line for a single treatment.

simple regression line The slope is the sample covariance of the treatment and outcome divided by the sample variance of the treatment. The intercept is the mean value of the outcome minus the slope times the mean value of the treatment.

simultaneity This can arise when one or more of the treatments is determined at the same time as the outcome; often occurs when some amount of reverse causality occurs.

standard deviation The square root of the variance.

strength (of an inductive argument) The degree of confidence in the conclusion.

structured data Data with well-defined units of observation for which corresponding information is identifiable; they are the data that come in a spreadsheet format.

subjective degrees of support Degrees of support based on opinion and lacking a statistical foundation.

sum of squared residuals (SS_{Res}) The sum of the squared residuals.

T

t-statistic (t -stat) The difference between the sample mean and the hypothesized population mean ($\bar{X} - K$) divided by the sample standard deviation (S/\sqrt{N}).

test statistic Any single value derived from a sample that can be used to perform a hypothesis test.

time-series data Data that exhibit only variation in time.

total sum of squares (TSS) The sum of the squared difference between each observation of Y and the average value for Y .

transposition Any time a group of assumptions implies a conclusion (A implies B), then it is also true that any time the conclusion

339

does not hold (not B), at least one of the assumptions must not hold (not A).

treatment Something that is administered to members of at least one participating group.

treatment effect The change in the outcome resulting from variation in the treatment.

two stage least squares (2SLS) regression The process of using two regressions to measure the causal effect of a variable while utilizing an instrumental variable.

U

unbiased estimator An estimator whose mean is equal to the population parameter it is used to estimate.

unconditional correlation The standard measure of correlation.

unit of observation The entity for which information has been collected.

unstructured data Any data that cannot be classified as structured.

V

variance A common measure for the spread of a distribution.

variance inflation factor (VIF) For an independent variable—say, X_1 —is equal to $\frac{1}{1-R_{X_i}^2}$, where $R_{X_1}^2$ is the R -squared from regressing that independent variable (X_1) on all other independent variables (X_2, \dots, X_K) for a given determining function.

W

weak instrument An instrumental variable that has little partial correlation

with the endogenous variable whose causal effect on an outcome it is meant to help measure.

within estimation Uses regression analysis of within-group differences in variables to estimate the parameters in the fixed effects data-generating process, except for those corresponding to the fixed effects (and the constant).

INDEX

A

- Active prediction
 - background music in retail stores, 25
 - business strategy formation, 25–26
 - causal regression analysis, 179–181
 - deductive/inductive reasoning, 77–78
 - defined, 24
 - explanations, 326–331
 - R -squared, 183
- Ad duration and clicks, 329–331
- Airlines’ on-time performance and multimarket contact, 253–254
- Alibi, 39
- Alternative data populations, 296
- Applications (real-world problems), 322–334
 - ad duration and clicks, 329–331
 - auto ad budget and revenues, 332–333
 - data-driven conclusions, 323–326
 - explanations for data analysis and active predictions, 326–331
 - grocery store price promotions, 325–326
 - insurance claims and deductibles, 326–328
 - machine maintenance and quality, 333–334
 - projects, 331–334
 - switching insurance, 324–325
 - tablet price and profits, 331–332
 - tennis analytics, 323–324

wearable features and sales, 328–329
Aristotle, 34
Aslin, Richard, 105
Association analysis, 15, 16
ATE. See Average treatment effect (ATE)
Auto ad budget and revenues, 332–333
Availability bias, 49
Average treatment effect (ATE), 90

B

Banner ads, 85, 95
Base group, 205
Baseball queries, 13, 14
Bias
 availability, 49
 confirmation, 49
 omitted variable, 314–317
 predictable-world, 49
 selection. See Selection bias
Binary dependent variable, 261
Bottom-up logic, 43. See also Inductive reasoning
Broadband availability, 238
Business analytics, 3
Business strategy, 4
Business strategy formation, 25–26

C

Campaign evaluation, 19
Card, David, 243
Categorical variable, 204
Causal inference. See also Causality
 advanced methods, 224–257

assumptions, 188–199
business applications, 18–19
campaign evaluation, 19
confounding factors, 196
control variables, 200–209
defined, 17
determining function. See Determining function
direct/indirect causal relationship, 18
endogeneity. See Endogeneity problem
form of determining function, 213–218
fundamental approaches, 187–223
instrumental variable. See Instrumental variable
nonrandom samples, 193–196
panel data methods. See Panel data methods
prediction, 19
proxy variable, 210–212
scientific method, 89–94

Causality. See also Causal inference
correlation, contrasted, 156
experiments vs. causal regression analysis, 180
lack of correlation between “other factors,” 184
regression analysis, 170–182

Causality model, 172

Center
distribution, 60
sample, 61

Changing the offer to change the odds, 258–259, 287–288

Chapter-opening challenges. See dataCHALLENGE

Classification, 15

Click-through rate, 85, 95

Cluster analysis, 15, 16, 16f

CNN, 50

Coin flips, 46–47

Collector selection bias, 49
Confidence interval
correlational regression analysis, 166–167

341

deductive reasoning, 68
defined, 64
determining function, 176
imperfect multicollinearity, 308
90%, 66, 67
95%, 64, 66, 67
99%, 66, 67
probit and logit models, 282
scientific method, 99–103
treatment effect, 103
Confirmation bias, 49
Confounding factors, 196
Consistency of regression estimators
determining functions, 175
population correlations, 164
Consistent estimator, 163
Constructing a representative sample, 190–192
Continuous random variable, 57, 61
Control variables, 200–209
Cord cutting, 262
Correlation
causality, contrasted, 156
partial, 156, 159
positively correlated, 156
regression analysis, 157–170
sample, 156
semi-partial, 159
unconditional, 158
Correlation model, 172

Correlation vs. causality, 156
Cost variables, 237
Cross-sectional data, 7–8, 8t, 10
Customer testimonies, 45
Cutoff values
p-values, 75
standard deviations, 66, 67

D

Dancers with store signs outside restaurant, 83, 109
Dashboard, 20, 21f
Data
 cross-sectional, 7–8, 8t, 10
 defined, 3
 experimental, 86
 nonexperimental, 104–108
 panel, 9, 9t, 10
 singular/plural, 4
 structured, 5–6, 6t
 time-series, 8, 9t, 10, 304
 unstructured, 6, 6t
Data analysis, 95–104, 326–331
Data-driven conclusions, 323–326
Data dump, 1, 26–27
Data gap, 297
Data-generating process (DGP)
 broadband availability, 238
 causal analysis, 154
 causality model, 172
 causality vs. correlation, 156
 controls, 200
 defined, 9, 153
 formal model, 11

general formulation, 153
informal concept, 10–11
linear probability model, 267
obesity and fatty milk, 181
OnlineEd example, 154
outcome Y , 174, 188
population regression equation, contrasted, 173
profits, promotion, and financing, 313
relationship between profits and price, 292
tax hike on profits, 239
where to park your truck (revenue), 183
within estimation, 251

Data mining, 15, 23

Data sample, 43

Data sanity check for a regression, 207

Data types

- cross-sectional data, 7–8, 8t, 10
- panel data, 9, 9t, 10
- pooled cross-sectional data, 8, 8t, 10
- time-series data, 8, 9t, 10

Database, 3

dataCHALLENGE

- changing the offer to change the odds, 258–259, 287–288
- dancers with store signs outside restaurant, 83, 109
- data dump, 1, 26–27
- knowing all customers by observing a few, 55, 79
- parking food truck in college town, 113, 145, 151, 183–184
- relationship between profits and price, 292–293, 318–319
- sex imbalance, 32, 51–52
- TV ads and web traffic, 224, 254–255
- working out at work, 187, 218–219

Datum, 4

Decision trees, 24

Deductive reasoning

- active predictions, 77–78
- concreteness, 40
- confidence interval, 68, 166–167
- confidence interval (correlational regression analysis), 166
- confidence interval (determining function), 176
- confidence interval (treatment effect), 103
- defined, 35
- direct proof, 35–36, 38
- disagreement about a conclusion, 40
- hypothesis testing, 76
- hypothesis testing (correlational regression analysis), 167
- hypothesis testing (determining function), 176–177
- hypothesis testing (treatment effect), 100
- inductive reasoning, contrasted, 43
- legal applications, 39
- robustness, 40–41
- schematic illustration, 35f
- transposition, 36–37, 38

Deductive reasoning process, 35f

Degree of support, 44

Descriptive statistics, 13

Determining function

- confidence interval, 176
- consistency of regression estimator, 175
- defined, 153
- form of, 213–218
- hypothesis testing, 176–177
- linear functional form, 213
- log functional form, 217–218
- multiple treatments, 155
- probit/logit models, 281–283

quadratic functional form, 213

Deterministic variable, 57

DGP. See Data-generating process (DGP)

Dichotomous dependent variable, 261–263, 272t. See also Prediction for a dichotomous variable

Dichotomous treatment, 117

Difference-in-differences (diff-in-diff), 239–242

- defined, 241
- endogeneity problem, 242
- minimum wage, 243

Direct causal relationship, 18

Direct proof, 35–36, 38

Disagreement about a conclusion, 40

Discovery of penicillin, 88

Discrete random variable, 57, 61

Distribution of random variables, 57–61

Distribution of sample mean for hypothesized population mean, 71–72

Dummy variable, 202–206, 306

Dummy variable estimation, 246–247

E

Econometric models, 18

Economic climate, 212

Effect of the treatment on the treated (ETT), 92

Elasticity, 217

Empirically testable conclusion, 41–42

Employment offers and prospective employee acceptances, 258–259, 287–288

Endogeneity problem, 196–199

- control variables, 200
- defined, 196
- diff-in-diff, 240, 242
- fixed-effects model, 246, 251–252
- forms, 199, 314

identification problem, 309, 312–313
linear probability model, 270
measurement error, 199
omitted variable, 199
probit and logit models, 286
proxy variable, 210–212
signing omitted variable bias, 314–317
simultaneity, 199
variable co-movement, 309, 312–313

Error term, 172

Estimator, 63

ETT. See Effect of the treatment on the treated (ETT)

Evaluating assumptions, 45–48

Excel. See Microsoft Excel

Exogenous and relevant, 227–228, 234–236

Exogenous as an instrumental variable, 227–228

Exogenously altered, 23

Expected value, 60

Experience goods, 45

Experiment, 86

Experimental data, 86

Extrapolation and interpolation, 297–303

- changing the population, 299–300
- definitions, 297
- functional form assumption, 300–302
- identification problems, 298–299
- remedyng the identification problems, 299–302
- time-series data, 304

F

Facebook, 180

Fatty milk and obesity, 181
“Fit” of regression equation, 182
Fixed effects, 245
Fixed-effects model, 242–253
 basic implications (Reasoning Box 8.3), 252
 comparing estimation methods, 250–252
 controls beyond fixed effects and time dummies, 246
 controls for the groups, 246
 data-generating process, 245
 defined, 245
 dummy variable estimation, 246–247
 endogeneity problem, 246, 251–252
 within estimation, 248–250, 251
 group switching, 250
 grouping process, 252–253
 R-squared, 251
 time controls, 245–246
Fleming, Alexander, 88
Four “W” questions, 7
Fox News, 50
Fraud detection, 3
Functional form
 determining function (causal inference), 213–218
 extrapolation and interpolation, 300–302

G

Galton, Francis, 132
GDP. See Gross domestic product (GDP)
Generalized method of moments (GMM), 233
GMM. See Generalized method of moments (GMM)
Good control, 201
Good proxy variable, 211–212
Google search, 12

Grocery store price promotions, 325–326

Gross domestic product (GDP), 212

Guilt and innocence, 39

H

Harper, Bryce, 14

Healey, Ben, 95

Homoscedasticity, 166

Hypothesis, 86

Hypothesis test, 70

Hypothesis testing, 70–76

correlational regression analysis, 167

deductive reasoning, 76

determining function, 176–177

inductive reasoning, 76

p-value, 74

probit and logit models, 282–283

scientific method, 96–99, 100

treatment effect, 100

I

Identification and data assessment, 292–321

endogeneity, 309, 312–313

extrapolation and interpolation, 297–303

identified, defined, 294

signing omitted variable bias, 314–317

variable co-movement. See Variable co-movement

Identified, 294

i.i.d. See Independent and identically distributed (i.i.d.)

Imperfect multicollinearity

defined, 304–305

identification problems, 307–309

profits, promotion, and financing, 313
remedies, 312
variance inflation factor (VIF), 308–309, 313

Inconsistent estimator, 309f

Independent, 64

Independent and identically distributed (i.i.d.), 64

Indirect causal relationship, 18

Inductive probability, 44

Inductive reasoning

- active predictions, 77–78
- basic form, 43
- confidence interval (correlational regression analysis), 167
- confidence interval (determining function), 176
- confidence interval (probit/logit determining function), 282
- confidence interval (treatment effect), 103
- customer testimonies, 45
- deductive reasoning, contrasted, 43
- defined, 43
- degree of support, 44
- evaluating assumptions, 45–48
- hypothesis testing, 76
- hypothesis testing (correlational regression analysis), 167
- hypothesis testing (determining function), 177
- hypothesis testing (probit/logit determining function), 283
- hypothesis testing (treatment effect), 100
- making general statement based on specific observations, 43
- population parameters, 64
- selection bias, 50
- strength, 44

344

Inspirational speeches, 40

Instrumental variable, 225–238

- business applications, 237–238

cost variables, 237
defined, 226
exogenous and relevant, 227–228, 234–236
policy change, 238
two-stage least squares regression (2SLS), 228–233
weak instrument, 236
Insurance claims and deductibles, 326–328
Intercept, simple regression line, 129
Interpolation, 297. See also Extrapolation and interpolation
Interval variable, 204
Irrelevant variable, 208
Ivy League degree and marginal revenue product of employees, 317–318

K

Key performance indicators (KPIs), 19
Kidd, Celeste, 105
Knockoff purses, 18
Knowing all customers by observing a few, 55, 79
KPIs. See Key performance indicators (KPIs)
Krueger, Alan, 243

L

LAD. See Least absolute deviations (LAD)
Laffer curve, 215–216
Lag information, 19, 20
Latent variable, 271, 272t
Lead information, 20, 22
Least absolute deviations (LAD), 133, 134
Least squares
 ordinary least squares (OLS), 133–135
 two-stage least squares regression (2SLS), 228–233
Lees, Gavin, 95

Level-log, 217, 218t
Limit-violating prediction, 268–269
Limited dependent variable, 260–263
Linear determining function, 213
Linear probability model, 263–270
 data-generating process, 267
 defined, 263
 endogeneity problem, 270
 interpretation, 266
 limit-violating prediction, 268–269
 merits/advantages, 267
 probit and logit models, 286
 shortcomings/limitations, 267–269, 285–286
Linear regression, 15, 142–144
Local labor laws, 238
Local sales tax, 238
Log functional form, 217–218
Log-level, 217, 218t
Log-log, 217, 218t
Logic, 34
Logistic distribution, 274
Logit model, 274. See also Probit and logit models

M

Machine maintenance and quality, 333–334
Margin of error, 70
Marginal effects, 276–278, 286
Marijuana and educational attainment, 203
Marshmallow test (marshmallows and reliability), 105
Maximum likelihood estimation (MLE), 279–281
Mayer, Marissa, 77
Measurement error, 199
Mehr, Samuel, 102

Member selection bias, 50

Method of linear least squares, 133

Microsoft Excel

- norm.dist, 59
- NORM.S.DIST(-1, TRUE), 277
- p*-value, 75
- perfect multicollinearity, 306–307, 307t
- regression output, 138t

Microsoft Surface, 47

Minimum wage, 238, 243

MLE. See Maximum likelihood estimation (MLE)

Model fit, 24

Moment conditions, 115, 278. See also Sample moment

MSNBC, 50

Multi-level treatment, 124

Multichannel video programming distributors (MVPDs), 262

Multicollinearity. See Imperfect multicollinearity; Perfect multicollinearity

Multimarket contact and airlines' on-time performance, 253–254

Multiple regression, 140

Multiple regression hyperplane, 140

Multiple regression plane, 140

Music

- background music in retail stores, 25
- music training and intelligence, 102

MVPDs. See Multichannel video programming distributors (MVPDs)

N

Neural networks, 24

News and sensationalism, 49–50

News network polls, 50

90% confidence interval, 66, 67

95% confidence interval, 64, 66, 67
99% confidence interval, 66, 67
Non-zero average treatment effect, 92
Nonexperimental data, 104–108
Nonrandom samples, 193–196
Normal random variable, 58, 58f
norm.dist, 59
NORM.S.DIST(-1, TRUE), 277
Null hypothesis, 71

O

Obesity and fatty milk, 181
Objective degree of support, 44, 57
Objective function, 133
OLS. See Ordinary least squares (OLS)
OLS multiple regression, 140
OLS multiple regression (hyper)plane, 140
Omitted variable, 199
Omitted variable bias, 314–317
Ordinal variable, 204
Ordinary least squares (OLS), 133–135
“Other factors,” 153, 174, 184, 196
Outcome probabilities, 42
Outlier detection, 15, 16, 16f

P

p-value, 74, 74f, 75, 308
Palmeri, Holly, 105
Panel data, 9, 9t, 10, 239
Panel data methods, 239–253
 business applications, 252–253
 difference-in-differences (diff-in-diff), 239–242

fixed-effects model. See Fixed-effects model

grouping process, 252–253

types of panel data, 252

Parking food truck in college town, 113, 145, 151, 183–184

Partial correlation, 156, 159

Passive prediction

- applications in business, 24
- correlational regression analysis, 169–170
- defined, 23
- model fit, 24, 182–183
- neural networks/decision trees, 24
- political races, 25
- R-squared, 183
- weather forecasting, 23–24

Pattern, 15

Pattern discovery, 15–17, 23

Penicillium notatum, 88

Perfect multicollinearity

- defined, 304–305
- dummy variable, 306
- identification problems, 305–307
- remedies to identification problems, 310–312

Physical fitness and academic success, 157

Pivot table, 13, 14t

Policy change, 238

Political polls, 25, 70

Polynomial, 214

Pooled cross-sectional data, 8, 8t, 10

Population, 43

Population mean, 60

Population parameter, 57, 60

Population regression equation, 163, 169, 172–175

Positively correlated, 156

Predictable-world bias, 49

Prediction, 19

Prediction for a dichotomous variable, 258–291

- dichotomous dependent variable, 261–263, 272t
- limited dependent variable, 260–263
- linear probability model. See Linear probability model
- model selection (choosing the “right” model), 270, 286–287
- probit/logit models. See Probit and logit models

Predictive analytics, 22–24

- active prediction. See Active prediction
- business strategy, 4–5, 25–26
- defined, 3
- models, 5
- passive prediction. See Passive prediction

Predictive analytics models, 5

Price regulations, 238

Price variation, 227f

Price vs. profit (dataCHALLENGE), 292–293, 318–319

Prince, Jeffrey, 253

Probability density function (pdf), 57

Probability function, 57

346

Probit and logit models, 270–286

- assumptions, 280–281
- complexity of calculating marginal effects, 286
- confidence intervals, 282
- endogeneity, 286
- estimation and interpretation, 278–284
- hypothesis testing, 282–283
- instrumental variables and/or fixed effects, 286
- latent variable, 271, 272t
- logit model, defined, 274
- marginal effects, 276–278, 286

maximum likelihood estimation (MLE), [279–281](#)
merits and shortcomings, [285–286](#)
overcoming shortcomings of linear probability model, [286](#)
probit/logit model, compared, [275](#)
probit model, defined, [273](#)
STATA “logit” command, [284](#), [284t](#)
STATA “probit” command, [283](#), [284t](#)
Profits, promotion, and financing, [313](#)
Projecting trends, [304](#)
Projects
 auto ad budget and revenues, [332–333](#)
 machine maintenance and quality, [333–334](#)
 tablet price and profits, [331–332](#)
Proxy variable, [210–212](#)
Purse knockoffs, [18](#)

Q

Quadratic determining function, [213](#)
Quadratic relationship, [214f](#)
Quarter example (flipping a quarter), [46–47](#)
Query, [12–14](#)
Query software, [12](#)

R

R-squared
 active prediction, [183](#)
 defined, [182](#)
 fixed-effects model, [251](#)
 passive prediction, [183](#)
Random sample, [63](#), [189–190](#)
Random variable, [57](#)
Ratings, [144](#)

Real-world problems. See Applications (real-world problems)

Reasoning

deductive. See Deductive reasoning

defined, 34

inductive. See Inductive reasoning

Regression analysis, 113–186

causality, 170–182

correlation, 157–170

data-generating process. See Data-generating process (DGP)

data sanity check, 207

defined, 117

determining function. See Determining function

experiments vs. causal regression analysis, 180

“fit” of regression equation, 182

least absolute deviations (LAD), 133, 134

least squares, 133–135

linear regression, 142–144

multiple regression, 140

multiple regression plane/hyperplane, 140

multiple treatments, 135–141

ordinary least squares (OLS), 133–135

“other factors,” 153, 174, 184

population regression equation, 163, 169, 172–175

R-squared, 182–183

ratings, 144

regression line (dichotomous treatment), 116–123

regression line (multi-level treatment), 124–132

regression plane/hyperplane, 138, 139

sample moment equations, 133–135

simple regression, 140

simple regression line, 129–132

two-stage least squares regression (2SLS), 228–233

Regression hyperplane, 139

Regression line

dichotomous treatment, [116–123](#)

full population, [164f](#)

multi-level treatment, [124–132](#)

rocking chair example, [301f](#)

Regression line for a dichotomous treatment, [119, 122](#)

Regression plane, [138, 138f, 139](#)

Relevant as an instrumental variable, [228, 235–236](#)

Report, [20](#)

Representative sample, [190–192](#)

Residual, [121](#)

Retail stores, background music, [25](#)

Risk management, [3](#)

Robustness, [40–41](#)

Rude sales clerks and sales volume, [108](#)

S

Salary offers and prospective employee acceptances, [258–259, 287–288](#)

Sales tax, [238](#)

Sample correlation, [156](#)

Sample covariance, [156](#)

347

Sample mean, [61](#)

Sample moment, [133](#)

Sample moment equations, [133–135, 278](#)

Sample of size N , [61](#)

Sample standard deviation, [61](#)

Sample statistics, [61](#)

Sample variance, [61](#)

Scatterplot, [17f](#)

Scientific method, [83–112](#)

analysis and conclusions, [86–87](#)

asking a question, 85
average treatment effect (ATE), 90
background research, 86
causal inference, 89–94
confidence interval, 99–103
data analysis, 95–104
defined, 84
discovery of penicillin, 88
effect of the treatment on the treated (ETT), 92
experiment, 86
experimental vs. non-experimental data, 86, 104–108
formulating a hypothesis, 86
hypothesis testing, 96–99, 100
overview, 88t
reporting the findings, 87
steps in process, 85f
uses, 87

Scorecard, 21, 22f

Selected sample, 193–196

Selection bias

- collector, 49
- defined, 49, 92
- inductive reasoning, 50
- member, 50
- news network polls, 51
- treatment effect, 93, 94

Semi-partial correlation, 159, 308

Sensationalism in the news, 49–50

Sex imbalance, 32, 51–52

Signing omitted variable bias, 314–317

Simon, Daniel, 253

Simple regression, 140

Simple regression line, 129–132

Simultaneity, 199
Situational batting averages, 14
Slope, simple regression line, 129
Spread
 distribution, 60
 sample, 61
 SS_{Res} . See Sum of squared residuals (SS_{Res})
Standard deviation, 60
STATA
 “logit” command, 284, 284t
 “probit” command, 283, 284t
Strategies vs. tactics, 4n
Strength, 44
Structured data, 5–6, 6t
Stylized examples, 322. See also Applications (real-world problems)
Subjective degrees of support, 44
Sum of squared residuals (SS_{Res}), 182
Switching insurance, 324–325

T

t-distribution, 67–68, 75
t-stat, 73, 75
Tablet price and profits, 331–332
Tactics vs. strategies, 4n
Target, 24
Telecommuting (working from home), 77
Tennis analytics, 323–324
Test statistic, 73
Time-series data, 8, 9t, 10, 304
Top-down logic, 35. See also Deductive reasoning
Total sum of squares (TSS), 182
Transposition, 36–37, 38
Treatment, 86

Treatment effect, 86
Trout, Mike, 14
TSS. See Total sum of squares (TSS)
TV ads and web traffic, 224, 254–255
Two-stage least squares regression (2SLS), 228–233

U

Unbiased estimator, 65
Unconditional correlation, 158
Unit of observation, 5, 6–7. See also Data types
Unstructured data, 6, 6f
Uses of data analysis
 causal inference, 17–19
 pattern discovery, 15–17
 queries, 12–14

V

Variable
 categorical, 204
 continuous random, 57, 61
 control, 200–209
348

 cost, 237
 deterministic, 57
 dichotomous dependent, 261–263, 272t
 discrete random, 57, 61
 dummy, 202–206
 instrumental. See Instrumental variable
 interval, 204
 irrelevant, 208
 latent, 271, 272t
 limited dependent, 260–263

normal random, 58, 58f
omitted, 199
ordinal, 204
proxy, 210–212
random, 57
Variable co-movement, 303–317
 endogeneity, 309, 312–313
 identification problems, 305–310
imperfect multicollinearity. See Imperfect multicollinearity
perfect multicollinearity. See Perfect multicollinearity
remedies to identification problems, 310–313
signing omitted variable bias, 314–317
variance inflation factor (VIF), 308–309, 313
Variance, 60
Variance inflation factor (VIF), 308–309, 313
VIF. See Variance inflation factor (VIF)

W

“W” questions, 7
Weak instrument, 236
Wearable features and sales, 328–329
Weather forecasting, 23–24
Weierstrass theorem, 214–215
Within estimation, 248–250, 251
Working from home (telecommuting), 77
Working out at work, 187, 218–219

Z

Z score, 284t