Real Econometrics

The Right Tools to Answer Important Questions

Real Econometrics

The Right Tools to Answer Important Questions

Second Edition

Michael A. Bailey

New York Oxford
OXFORD UNIVERSITY PRESS

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide. Oxford is a registered trade mark of
Oxford University Press in the UK and certain other countries.

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America.

For titles covered by Section 112 of the US Higher Education Opportunity Act, please visit www.oup.com/us/he for the latest information about pricing and alternate formats.

Library of Congress Cataloging-in-Publication Data

Names: Bailey, Michael A., 1969- author.
Title: Real econometrics : the right tools to answer important questions / Michael A. Bailey.
Description: Second Edition. | New York : Oxford University Press, [2019] | Revised edition of the author's Real econometrics, [2017] | Includes bibliographical references and index.
Identifiers: LCCN 2018046855 (print) | LCCN 2018051766 (ebook) | ISBN 9780190857486 (ebook) | ISBN 9780190857462 (pbk.) | ISBN 9780190857523 (looseleaf)
Subjects: LCSH: Econometrics–Textbooks. | Economics–Study and teaching.
Classification: LCC HB139 (ebook) | LCC HB139 .B344 2019 (print) | DDC 330.01/5195–dc23
LC record available at https://lccn.loc.gov/2018046855

Printing number: 9 8 7 6 5 4 3 2 1

Printed in the United States of America

CONTENTS

APPENDICES

LIST OF FIGURES

LIST OF TABLES

USEFUL COMMANDS FOR STATA

Task	Command	Example	Chapter
Help	help	help summarize	2
Comment line	*	* This is a comment line	2
Comment on command line	/* */	use "C:\Data.dta" /* This is a comment */	
Continue line	/* */	reg y X1 X2 X3 /* */ X4 X5	2
Load Stata data file	use	use "C:/Data.dta"	2
Load text data file	insheet	insheet using "C:/Data.txt"	2
Display variables in memory	list	list /* Lists all observations for all variables */	2
		list Y X /* Lists all observations for Y and X */	2
		list X in 1/10 /* Lists first 10 observations for X */	2
Descriptive statistics	summarize	summarize X1 X2 Y	2
Frequency table	tabulate	tabulate X1	2
Scatter plot	scatter	scatter Y X	2
		scatter Y X, mlabel(name) /* Adds labels */	2
Limit data	if	summarize X1 if X2 > 1	2
Equal (as used in if statement, for example)	==	summarize X1 if X2 == 1	2
Not equal	!=	summarize X1 if X2!=0	2
And	&	list X1 if X2 == 1 & X3 > 18	2
Or	|	list X1 if X2 == 1 | X3 > 18	2
Delete a variable	drop	drop X7	2
Missing data in Stata	.	* Caution: Stata treats missing data as having infinite value, so list X1 if X2 > 0 will include values of X1 for which X2 is missing	2
Regression	reg	reg Y X1 X2	3
Heteroscedasticity robust regression	, robust	reg Y X1 X2, robust	3
Generate predicted values	predict	predict FittedY /* Run this after reg command */	3
Add regression line to scatter plot	twoway, lfit	twoway (scatter Y X) (lfit Y X)	3
Critical value for *t* distribution, two-sided	invttail	display invttail(120, .05/2) /* For model with 120 degrees of freedom and $\alpha = 0.05$; note that we divide α by 2 */	4
Critical value for *t* distribution, one-sided	invttail	display invttail(120, .05) /* For model with 120 degrees of freedom and $\alpha = 0.05$ */	4
Critical value for normal distribution, two-sided	invnormal	display invnormal(.975) /* For $\alpha = 0.05$, note that we divide α by 2 */	4
Critical value for normal distribution, one-sided	invnormal	display invnormal(.05)	4
Two-sided *p* values		[Reported in reg output]	4
One-sided *p* values	ttail	display 2*ttail(120, 1.69) /* For model with 120 degrees of freedom and a t statistic of 1.69 */	4
Confidence intervals		[Reported in reg output]	4
Produce standardized regression coefficients	, beta	reg Y X1 X2, beta	5
Produce standardized variable	egen	egen X_std = std(X) /* Creates variable called X_std */	5

Task	Command	Example	Chapter
F test	test	test X1 = X2 = 0 /* Run this after regression with X1 and X2 in model */	5
Critical value for F test	invF	display invF(2, 120, 0.95) /* Degrees of freedom equal 2 and 120 and $\alpha = 0.05$ */	5
p value for F statistic	Ftail	Ftail(2, 1846, 7.77) /* Degrees of freedom equal 2 and 1846 and F statistic = 7.77*/	5
Difference of means test using OLS	reg	reg Y Dum /* Where Dum is a dummy variable */	6
Create an interaction variable	gen	gen DumX = Dum * X	6
Include dummies for categorical variable	i.varname	reg Y i.X1 /* Includes appropriate number of dummy variables for categorical variable X1 */	6
Set reference category	ib#.varname	reg Y ib2.X1 /* Sets 2nd category as reference category */	6
Create a squared variable	gen	gen X_sq = X^2	7
Create a logged variable	gen	gen X_log =log(X)	7
Generate dummy variables for each unit	tabulate and generate	tabulate City, generate(City_dum)	8
LSDV model for panel data	reg	reg Y X1 X2 City_dum2 - City_dum80	8
De-meaned model for panel data	xtreg	xtreg Y X1 X2, fe i(City)	8
Two-way fixed effects	xtreg	xtreg Y X1 X2 i.year Yr2- Yr10, fe i(City)	8
2SLS model	ivregress	ivregress 2sls Y X2 X3 (X1 = Z), first	9
Probit	probit	probit Y X1 X2 X3	12
Normal CDF	normal	normal(0) /* The normal CDF evaluated at 0 (which is 0.5)*/	12
Logit	logit	logit Y X1 X2 X3	12
Critical value for χ^2 test	invchi2	display invchi2(1, 0.95) /* Degrees of freedom = 1 and 0.95 confidence level */	12
Account for autocorrelation in time series data	prais	tsset Year prais Y X1 X2, corc twostep	13
Include lagged dependent variable	L.Y	reg Y L.Y X1 X2 /* Run tsset command first */	13
Augmented Dickey-Fuller test	dfuller	dfuller Y, trend lags(1) regress	13
Generate draws from standard normal distribution	rnormal	gen Noise = rnormal(0,1) /* Length will be same as length of variables in memory */	14
Indicate to Stata unit and time variables	tsset	tsset ID time	15
Panel model with autocorrelation	xtregar	xtregar Y X1 X2, fe rhotype(regress) twostep	15
Include lagged dependent variable	L.Y	xtreg Y L.Y X1 X2, fe i(ID)	15
Random effects panel model	, re	xtreg Y X1 X2, re	15

USEFUL COMMANDS FOR R

Task	Command	Example	Chapter
Help	?	?mean # Describes the "mean" command	2
Comment line	#	# This is a comment	2
Load R data file	load	Data = load("C:/Data.RData")	2
Load text data file	read.table	Data = read.table("C:/Data.txt", header = TRUE)	2
Display names of variables in memory	objects	objects() # Will list names of all variables in memory	2
Display variables in memory	[enter variable name]	X1 # Display all values of this variable; enter directly in console or highlight in editor and press ctrl-r	2
		X1[1:10] # Display first 10 values of X1	2
Missing data in R	NA		
Mean	mean	mean(X1)	2
		mean(X1, na.rm=TRUE) # Necessary if there are missing values	
Variance	var	var(X1)	2
		var(X1, na.rm=TRUE) # Necessary if there are missing values	
		sqrt(var(X1)) # This is the standard deviation of X1	
Minimum	min	min(X1, na.rm=TRUE)	2
Maximum	max	max(X1, na.rm=TRUE)	2
Number of observations	sum and is.finite	sum(is.finite(X1))	2
Frequency table	table	table(X1)	2
Scatter plot	plot	plot(X, Y)	2
		text(X, Y, name) # Adds labels from variable called "name"	2
Limit data (similar to an if statement)	[]	plot(Y[X3<10], X1[X3<10])	2
Equal (as used in if statement, for example)	==	mean(X1[X2==1]) # Mean of X1 for cases where X2 equals 1	2
Not equal	!=	mean(X1[X1!=0]) # Mean of X1 for observations where X1 is not equal to 0	2
And	&	X1[X2 == 1 & X3 > 18]	2
Or	|	X1[X2 == 1 | X3 > 18]	2
Regression	lm	lm(Y ˜X1 + X2) # lm stands for "linear model"	3
		Results = lm(Y˜X) # Creates an object called "Results" that stores coefficients, standard errors, fitted values, and other information about this regression	3
Display results	summary	summary(Results) # Do this after creating "Results"	3
Install a package	install.packages	install.packages("AER") # Only do this once for each computer	3
Load a package	library	library(AER) # Include in every R session in which we use package specified in command	
Heteroscedasticity robust regression	coeftest	coeftest(Results, vcov = vcovHC(Results, type = "HC1")) # Need to install and load AER package for this command. Do this after creating OLS regression object called "Results"	3
Generate predicted values	$fitted.values	Results$fitted.values # Run after creating OLS regression object called "Results"	3
Add regression line to scatter plot	abline	abline(Results) # Run after plot command and after creating "Results" object based on a bivariate regression	3

Task	Command	Example	Chapter
Critical value for t distribution, two-sided	qt	qt(0.975, 120) # For $\alpha = 0.05$ and 120 degrees of freedom; divide α by 2	4
Critical value for t distribution, one-sided	qt	qt(0.95, 120) # For $\alpha = 0.05$ and 120 degrees of freedom	4
Critical value for normal distribution, two-sided	qnorm	qnorm(0.975) # For $\alpha = 0.05$; divide α by 2	4
Critical value for normal distribution, one-sided	qnorm	qnorm(0.95) # For $\alpha = 0.05$	4
Two-sided p values		[Reported in summary(Results) output]	
One-sided p values	pt	2*(1-pt(abs(1.69), 120)) # For model with 120 degrees of freedom and a t statistic of 1.69	4
Confidence intervals	confint	confint(Results, level = 0.95) # For OLS object "Results"	4
Produce standardized regression coefficients	scale	Res.std = lm(scale(Y) ˜scale(X1) + scale(X2))	5
Display R squared	$r.squared	summary(Results)$r.squared	5
Critical value for F test	qf	qf(.95, df1 = 2, df2 = 120) # Degrees of freedom equal 2 and 120 and $\alpha = 0.05$	5
p value for F statistic	pf	1 - pf(7.77, df1=2, df2=1,846) # For F statistic = 7.77, and degrees of freedom equal 2 and 1846	5
Include dummies for categorical variable	factor	lm(Y ∼ factor(X1)) # Includes appropriate number of dummy variables for categorical variable X1	6
Set reference category	relevel	X1 = relevel(X1, ref = “south”) # Sets 2nd category as reference category; include before OLS model	6
Difference of means test using OLS	lm	lm(Y˜Dum) # Where Dum is a dummy variable	6
Create an interaction variable		DumX = Dum * X # Or use <- in place of =	6
Create a squared variable		X_sq = Xˆ2	7
Create a logged variable		X_log =log(X)	7
LSDV model for panel data	factor	Results = lm(Y ∼ X1 + factor(country)) # Factor adds a dummy variable for every value of variable called country	8

One-way fixed-effects model (de-meaned)	plm	library(plm)
		Results = plm(Y ˜X1+ X2+ X3, data = dta, index=c("country"), model="within")
			8

Two-way fixed-effects model (de-meaned)	plm	library(plm)
		Results = plm(Y ˜X1+ X2+ X3, data = dta, index=c("country", "year"), model="within", effect = "twoways")
			8

2SLS model	ivreg	library(AER)
		ivreg(Y ˜X1 + X2 + X3 |Z1 + Z2 + X2 + X3)
			9

Probit	glm	glm(Y ˜X1 + X2, family = binomial(link ="probit"))	12
Normal CDF	pnorm	pnorm(0) # The normal CDF evaluated at 0 (which is 0.5)	12
Logit	glm	glm(Y ˜X1 + X2, family = binomial(link ="logit"))	12
Generate draws from standard normal distribution	rnorm	Noise = rnorm(500) # 500 draws from standard normal distribution	14
Panel model with autocorrelation		[See Computing Corner in Chapter 15]	15
Include lagged dependent variable	plm with lag(Y)	Results = plm(Y ˜lag(Y) + X1 + X2, data = dta, index = c("ID", "time"), effect = "twoways")	15
Random effects panel model	plm with "random"	Results = plm(Y ˜X1 + X2, data = dta, model = "random")	15

PREFACE FOR STUDENTS: HOW THIS BOOK CAN HELP YOU LEARN ECONOMETRICS

"Less dull than traditional texts."—Student A.H.

"It would have been immensely helpful for me to have a textbook like this in my classes throughout my college and graduate experience. It feels more like an interactive learning experience than simply reading equations and facts out of a book and being expected to absorb them."—Student S.A.

"I wish I had had this book when I was first exposed to the material—it would have saved a lot of time and hair-pulling . . ."—Student J.H.

"Material is easy to understand, hard to forget."—Student M.H.

This book introduces the econometric tools necessary to answer important questions. Do antipoverty programs work? Does unemployment affect inflation? Does campaign spending affect election outcomes? These and many more questions are not only interesting but also important to answer correctly if we want to support policies that are good for people, countries, and the world.

When using econometrics to answer such questions, we need always to remember a single big idea: correlation is not causation. Just because variable Y rises when variable X rises does not mean that variable X *causes* variable Y to rise. The essential goal is to figure out when we can say that changes in variable X will lead to changes in variable Y.

This book helps us learn how to identify causal relationships with three features seldom found in other econometrics textbooks. First, it focuses on the tools that economic researchers use most. These are the *real* econometric techniques that help us make reasonable claims about whether X causes Y, and by using these tools, we can produce analyses that others can respect. We'll get the most out of our data while recognizing the limits in what we can say or how confident we can be.

This emphasis on *real econometrics* means that we skip obscure econometric tools that *could* come up under certain conditions. Econometrics is too often complicated by books and teachers trying to do too much. This book shows that we can have a sophisticated understanding of statistical inference without having to catalog every method that our instructor had to learn as a student.

Second, this book works with a single unifying framework. We don't start over with each new concept; instead, we build around a core model. That means there is a single equation and a unifying set of assumptions that we poke, probe, and expand throughout the book. This approach reduces the learning costs of moving through the material and allows us to go back and revisit material. As with any skill, we probably won't fully understand any given technique the first time we see it. We have to work at it; we have to work *with* it. We'll get comfortable; we'll see connections. Then it will click. Whether the skill is jumping rope, typing, throwing a baseball, or analyzing data, we have to do things many times to get good at it. By sticking to a unifying framework, we have more chances to revisit what we have already learned. You'll also notice that I'm not afraid to repeat myself on the important stuff. Really, I'm not afraid to repeat myself.

Third, this book uses many examples from the policy, political, and economic worlds. So even if you do not care about "two-stage least squares" or "maximum likelihood" in and of themselves, you will see how understanding these techniques will affect what you think about education policy, trade policy, election outcomes, and many other interesting issues. The examples and case studies make it clear that the tools developed in this book are being used by contemporary applied economists who are actually making a difference with their empirical work.

Real Econometrics is meant to serve as the primary textbook in an introductory econometrics course or as a supplemental text providing more intuition and context in a more advanced econometric methods course. As more and more public policy and corporate decisions are based on statistical and econometric analysis, this book can also be used outside of course work. Econometrics has infiltrated into every area of our lives—from entertainment to sports (I no longer spit out my coffee when I come across an article on regression analysis of National Hockey League players)—and a working knowledge of basic econometric techniques can help anyone make better sense of the world around them.

What's in This Book?

The preparation necessary to use this book successfully is modest. We use basic algebra a fair bit, being careful to explain every step. You do not need calculus. We refer to calculus when useful, and the book certainly could be used by a course that works through some of the concepts using calculus. However, you can understand everything without knowing calculus.

We start with two introductory chapters. Chapter 1 lays out the central challenge in econometrics. This is the challenge of making probabilistic yet accurate claims about causal relations between variables. We present experiments as an ideal way to conduct research, but we also show how experiments in the real world are tricky and can't answer every question we care about. This chapter provides the "big picture" context for econometric analysis that is every bit as important as the specifics that follow.

Chapter 2 provides a practical foundation related to good econometric practices. In every econometric analysis, data meets software, and if we're not careful, we lose control. This chapter therefore seeks to teach good habits about documenting analysis and understanding data.

The five chapters of Part One constitute the heart of the book. They introduce ordinary least squares (OLS), also known as regression analysis. Chapter 3 introduces the most basic regression model, the bivariate OLS model. Chapter 4 shows how to use OLS to test hypotheses. Chapters 5 through 7 introduce the multivariate OLS model and applications. By the end of Part One, you will understand regression and be able to control for anything you can measure. You'll also be able to fit curves to data and assess whether the effects of some variables differ across groups, among other skills that will impress your friends.

Part Two introduces techniques that constitute the contemporary econometric toolkit. These are the techniques people use when they want to get published—or paid. These techniques build on multivariate OLS to give us a better chance of identifying causal relations between two variables. Chapter 8 covers a simple yet powerful way to control for many factors we can't measure directly. Chapter 9 covers instrumental variable techniques, which work if we can find a variable that affects our independent variable but not our dependent variable. Instrumental variable techniques are a bit funky, but they can be very useful for isolating causal effects. Chapter 10 covers randomized experiments. Although ideal in theory, in practice such experiments often raise a number of challenges we need to address. Chapter 11 covers regression discontinuity tools that can be used when we're studying the effect of variables that were allocated based on a fixed rule. For example, Medicare is available to people in the United States only when they turn 65, and admission to certain private schools depends on a test score exceeding some threshold. Focusing on policies that depend on such thresholds turns out to be a great context for conducting credible econometric analysis.

Part Three contains a single chapter (Chapter 12) that covers dichotomous dependent variable models. These are simply models in which the outcome we care about takes on two possible values. Examples and case studies include high school graduation (someone graduates or doesn't), unemployment (someone has a job or doesn't), and alliances (two countries sign an alliance treaty or don't). We show how to apply OLS to such models and then provide more elaborate models that address the deficiencies of OLS in this context.

Part Four supplements the book with additional useful material. Chapter 13 covers time series data. The first part of the chapter is a variation on OLS; the second part introduces dynamic models that differ from OLS models in important ways. Chapter 14 derives important OLS results and extends discussion on specific topics. Chapter 15 goes into greater detail on the vast literature on panel data, showing how the various strands fit together.

Chapter 16 concludes the book with tips on adopting the mind-set of an econometric realist. In fact, if you are looking for an overall understanding of the power and limits of statistics, you might want to read this chapter first—and then read it again once you've learned all the statistical concepts covered in the other chapters.

How to Use This Book

Real Econometrics is designed to help you master the material. Each section ends with a "Remember This" box that highlights the key points of that section. If you remember what's in each of these boxes, you'll have a great foundation in statistics. Key Terms are boldfaced where they are first introduced in the text, defined briefly in the margins, and defined again in the glossary at the end of the book.

Review Questions and Discussion Questions appear at the end of selected sections. I recommend using these. Answering questions helps us be realistic about whether we're truly on track. What we're fighting is something cognitive psychologists call the "illusion of explanatory depth." That's a fancy way of saying we don't always know as much as we think we do. By answering the Review Questions and Discussion Questions, we can see where we are. The Review Questions are more concrete and have specific answers, which are found at the end of the book. The Discussion Questions are more open-ended and encourage us to explore how the concepts apply to issues we care about. Once invested in this way, we're no longer doing econometrics for the sake of doing econometrics; instead, we're doing econometrics to help us learn about important issues.

And remember, learning is not only about answering questions: coming up with your own questions for your instructor or classmates or the dude next to you on the bus is a great way to learn. Doing so will help you formulate exactly what is unclear and will open the door to an exchange of ideas. Heck, maybe you'll make friends with the bus guy or, worst case, you'll see an empty seat open up next to you . . .

Finally, you may have noticed that this book is opinionated and a bit chatty. This is not the usual tone of econometrics books, but being chatty is not the same as being dumb. You'll see real material, with real equations and real research—sometimes accompanied by smart-ass asides that you may not see in other books. This approach makes the material more accessible and also reinforces the right mind-set: econometrics is not simply a set of mathematical equations; instead, econometrics provides a set of practical tools that curious people use to learn from the world. But don't let the tone fool you. This book is not *Econometrics for Dummies*; it's *Real Econometrics*. Learn the material, and you will be well on your way to using econometrics to answer important questions.

PREFACE FOR INSTRUCTORS: HOW TO HELP YOUR STUDENTS LEARN ECONOMETRICS

We econometrics teachers have high hopes for our students. We want them to understand how econometrics can shed light on important economic and policy questions. Sometimes they humor us with incredible insight. The heavens part; angels sing. We want that to happen daily. Sadly, a more common experience is seeing a furrowed brow of confusion and frustration. It's cloudy and rainy in that place.

It doesn't have to be this way. If we distill the material to the most critical concepts, we can inspire more insight and less brow-furrowing. Unfortunately, conventional statistics and econometrics books all too often manage to be too simple and too confusing at the same time. Many are too simple in that they provide a semester's worth of material that hardly gets past rudimentary ordinary least squares (OLS). Some are too confusing in that they get to OLS by way of going deep into the weeds of probability theory without showing students how econometrics can be useful and interesting.

Real Econometrics is predicated on the belief that we are most effective when we teach the tools we use. What we use are regression-based tools with an increasing focus on experiments and causal inference. If students can understand these fundamental concepts, they can legitimately participate in analytically sound conversations. They can produce analysis that is interesting—and believable! They can understand experiments and the sometimes subtle analysis required when experimental methods meet social scientific reality. They can appreciate that causal effects are hard to tease out with observational data and that standard errors estimated on crap coefficients, however complex, do no one any good. They can sniff out when others are being naive or cynical. It is only when we muck around too long in the weeds of less useful material that statistics becomes the quagmire students fear.

Hence this book seeks to be analytically sophisticated in a simple and relevant way. It focuses on tools actually used by real analysts. Nothing useless. No clutter. To do so, the book is guided by three principles: relevance, opportunity costs, and pedagogical efficiency.

Relevance

Relevance is a crucial first principle for successfully teaching econometrics in the social sciences. Every experienced instructor knows that most students care more about the real world than math. How do we get such students to engage with econometrics? One option is to cajole them to care more and work harder. We all know how well that works. A better option is to show them how a sophisticated understanding of statistical concepts helps them learn more about the topics that concern them. Think of a mother trying to get a child to commit to the training necessary to play competitive sports. She *could* start with a semester of theory. . . . No, that would be cruel. And counterproductive. Much better to let the child play and experience the joy of the sport. Then there will be time (and motivation!) to understand nuances. Thus every chapter is built around examples and case studies on topics students might actually care about—topics like violent crime in the United States (Chapter 2), global warming (Chapter 7), and the relationship between alcohol consumption and grades (Chapter 11).

Learning econometrics is not that different from learning anything else. We need to care to truly learn. Therefore this book takes advantage of a careful selection of material to spend more time on the real examples that students care about.

Opportunity Costs

Opportunity costs are, as we all tell our students, what we have to give up to do something. So, while some topic might be a perfectly respectable part of an econometric toolkit, we should include it only if it does not knock out something more important. The important stuff all too often gets shunted aside as we fill up the early part of students' analytical training with statistical knick-knacks, material "some people still use" or that students "might see."

Therefore this book goes quickly through descriptive statistics and doesn't cover χ^2 tests for two-way tables, weighted least squares, and other denizens of conventional statistics books. These concepts—and many, many more—are all perfectly legitimate. Some are covered elsewhere (descriptive statistics are covered in elementary schools these days). Others are valuable enough to rate inclusion here in an "advanced material" section for students and instructors who want to pursue these topics further. And others simply don't make the cut. Only by focusing the material can we get to the tools used by researchers today, tools such as panel data analysis, instrumental variables, and regression discontinuity. The core ideas behind these tools are not particularly difficult, but we need to *make* time to cover them.

Pedagogical Efficiency

Pedagogical efficiency refers to streamlining the learning process by using a single unified framework. Everything in this book builds from the standard regression model. Hypothesis testing, difference of means, and experiments can be—and often are—taught independently of regression. Causal inference is sometimes taught with potential outcomes notation. There is nothing intellectually wrong with these approaches. But is using them pedagogically efficient? If we teach these as stand-alone concepts we have to take time and, more important, student brain space to set up each separate approach. For students, this is really hard. Remember the furrowed brows? Students work incredibly hard to get their heads around difference of means and where to put degrees of freedom corrections and how to know if the means come from correlated groups or independent groups and what the equation is for each of these cases. Then BAM! Suddenly the professor is talking about residuals and squared deviations. The transition is old hat for us, but it can overwhelm students first learning the material. It is more efficient to teach the OLS framework and use that to cover difference of means, experiments, and the contemporary canon of econometric analysis, including panel data, instrumental variables, and regression discontinuity. Each tool builds from the same regression model. Students start from a comfortable place and can see the continuity that exists.

An important benefit of working with a single framework is that it allows students to revisit the core model repeatedly throughout the term. Despite the brilliance of our teaching, students rarely can put it all together with one pass through the material. I know I didn't when I was beginning. Students need to see the material a few times, work with it a bit, and then it will finally click. Imagine if sports were coached the way we do econometrics. A tennis coach who said "This week we'll cover forehands (and only forehands), next week backhands (and only backhands), and the week after that serves (and only serves)" would not be a tennis coach for long. Instead, coaches introduce material, practice, and then keep working on the fundamentals. Working with a common framework throughout makes it easier to build in mini-drills about fundamentals as new material is introduced.

Course Adoption

Real Econometrics is organized to work well in three different kinds of courses. First, it can be used in an introductory econometrics course that follows a semester of probability and statistics. In such a course, students should probably be able to move quickly through the early material and then pick up where they left off, typically with multivariate OLS.

Second, this book can be used with students who have not previously (or recently) studied statistics, either in a one-semester course covering Part One or a year-long course covering the whole book. Using this book as a first course avoids the "warehouse problem," which occurs when we treat students' statistical education as a warehouse, filling it up with tools first and accessing them only later. One challenge is that things rot in a warehouse. Another challenge is that instructors tend to hoard a bit, putting things in the warehouse "just in case" and creating clutter. And students find warehouse work achingly dull. Using this book in a first-semester course avoids the warehouse problem by going directly to interesting and useful material, providing students with a more just-in-time approach. For example, they see statistical distributions, but in the context of trying to solve a specific problem rather than as an abstract concept that will become useful later.

Finally, *Real Econometrics* can be used as a supplement in a more advanced econometrics course, providing intuition and context that sometimes gets lost in the more technical courses.

Real Econometrics is also designed to encourage two particularly useful pedagogical techniques. One is interweaving, the process of weaving material from previous lessons into later lessons. Numbered sections end with a "Remember This" box that summarizes key points. Connecting back to these points in later lessons is remarkably effective at getting the material into the active part of students' brains. The more we ask students about omitted variable bias or multicollinearity or properties of instruments (and in sometimes surprising contexts), the more they become able to actively apply the material on their own.

The second teaching technique is to use frequent low-stakes quizzes to convert students to active learners with less stress than the exams they will also be taking. These quizzes need not be hard. They just need to give students a chance to independently access and apply the material. Students can test themselves with the Review Questions at the end of many sections, as the answers to these questions are at the back of the book. It can also be useful for students to discuss or at least reflect on the Discussion Questions at the ends of many sections, as these enable students to connect the material to real world examples. Brown, Roediger, and McDaniel (2014) provide an excellent discussion of these and other teaching techniques.

Overview

The first two chapters of the book serve as introductory material and introduce the science of statistics. Chapter 1 lays out the theme of how important—and hard—it is to generate unbiased estimates. This is a good time to let students offer hypotheses about questions of the day, because these questions can help bring to life the subsequent material. Chapter 2, which introduces computer programs and good practices, is a confidence builder that gets students who are not already acclimated to statistical computing over the hurdle of using statistical software.

Part One covers core OLS material. Chapter 3 introduces bivariate OLS. Chapter 4 covers hypothesis testing, and Chapter 5 moves to multivariate OLS. Chapters 6 and 7 proceed to practical tasks such as use of dummy variables, logged variables, interactions, and F tests.

Part Two covers essential elements of the contemporary econometric toolkit, including panel data, instrumental variables, analysis of experiments, and regression discontinuity. Chapter 10, on experiments, uses instrumental variables. Chapters 8, 9, and 11 can be covered in any order, however, so instructors can pick and choose among these chapters as needed.

Part Three contains a single chapter (Chapter 12) on dichotomous dependent variables. It develops the linear probability model in the context of OLS and uses the probit and logit models to introduce students to maximum likelihood. Instructors can cover this chapter any time after Part One if dichotomous dependent variables play a major role in the course.

Part Four introduces some advanced material. Chapter 13 discusses time series models, introducing techniques to account for autocorrelation and to estimate dynamic time series models; this chapter can also be covered at any time following Part One. Chapter 14 offers derivations of the OLS model and additional material on omitted variable bias. Instructors seeking to expose students to derivations and extensions of the core OLS material can use this chapter as an auxiliary to Chapters 3 through 5. Chapter 15 introduces more advanced topics in panel data. This chapter builds on material from Chapters 8 and 13.

Chapter 16 concludes the book by discussing ways to maximize the chances that we use econometrics properly to answer important questions about the world.

Every chapter ends with a series of learning tools. Each conclusion summarizes the learning objectives by section and provides a list of key terms introduced in the chapter (along with the page where first introduced). Each Further Reading section guides students to additional resources on the material covered in the chapter. The Computing Corners provide a guide to the syntax needed to implement the analysis discussed in the chapters. We provide this syntax for both Stata and R computing languages. Finally, the Exercises provide a variety of opportunities for students to analyze real data sets from important papers on interesting topics.

Several appendices provide supporting material. An appendix on math and probability covers background ranging from mathematical functions to important concepts in probability. In addition, citations and additional notes are linked to the text by page numbers and elaborate on some finer points. Answers to Review Questions are also provided.

Teaching econometrics is difficult. When the going gets tough it is tempting to blame students, to say they are unwilling to do the work. Before we go that route, we should recognize that many students find the material quite foreign and (unfortunately) irrelevant. If we can streamline what we teach and connect it to things students care about, we can improve our chances of getting students to understand the material, which not only is intrinsically interesting but also forms the foundation for all empirical work. When students understand, teaching becomes easier. And better. The goal of this book is to help get us there.

Supplements Accompanying *Real Econometrics*

A broad array of instructor and student resources for *Real Econometrics* are available online at www.oup.com/us/bailey.

Data

Much of the supplementary material for *Real Econometrics* focuses on data—through online access to the data sets referenced in the chapters, their documentation, and additional data sets. These include:

- Chapter-specific libraries of downloadable figures, graphs, and data sets (and their documentation) for the examples and exercises found in the text.

- Links to other data sets (both experimental and non-experimental) for creating new assignments.

Instructor's Manual

Each chapter in the Instructor's Manual provides an overview of the chapter goals and section-by-section teaching tips along with suggested responses to the in-chapter Discussion Questions. The Instructor's Manual also contains sample data sets for the Computing Corner activities and solutions to the Exercises found at the end of each chapter.

PowerPoint Presentations

Presentation slides offer bullet-point summaries as well as all the tables and graphs from the book to help guide and design lectures. A separate set of slides containing only the text tables and graphs is also available.

Computerized Test Bank

The computerized test bank that accompanies this text enables instructors to easily create quizzes and exams, using any combination of publisher-provided questions and their own questions. Questions can be edited and easily assembled into assessments that can then be exported for use in learning management systems or printed for paper-based assessments.

Learning Management Systems Support

For instructors using an online learning management system (e.g., Moodle, Sakai, Blackboard, or others), Oxford University Press can provide all the electronic components of the package in a format suitable for easy upload. Adopting instructors should contact their local Oxford University Press sales representative or OUP's Customer Service (800-445-9714) for more information.

ACKNOWLEDGMENTS

This book has benefited from close reading and probing questions from a large number of people, including students at the McCourt School of Public Policy at Georgetown University and my current and former colleagues and students at Georgetown, including Shirley Adelstein, Rachel Blum, David Buckley, Ian Gale, Ariya Hagh, Carolyn Hill, Mark Hines, Dan Hopkins, Jeremy Horowitz, Huade Huo, Wes Joe, Karin Kitchens, Jon Ladd, Jens Ludwig, Paasha Mahdavi, Jean Mitchell, Paul Musgrave, Sheeva Nesva, Hans Noel, Irfan Nooruddin, Ji Yeon Park, Parina Patel, Betsy Pearl, Lindsay Pettingill, Carlo Prato, Barbara Schone, George Shambaugh, Dennis Quinn, Chris Schorr, Frank Vella, and Erik Voeten. Credit (and/or blame) for the *Simpsons* figure goes to Paul Musgrave.

Participants at a seminar on the book at the University of Maryland, especially Antoine Banks, Brandon Bartels, Kanisha Bond, Ernesto Calvo, Sarah Croco, Michael Hanmer, Danny Hayes, Eric Lawrence, Irwin Morris, and John Sides, gave excellent early feedback.

In addition, colleagues across the country have been incredibly helpful, especially Allison Carnegie, Craig Volden, Sarah Croco, and Wendy Tam-Cho. Reviewers for Oxford University Press and other commentators have provided supportive yet probing feedback. These individuals include:

Steve Balla, *George Washington University*; Yong Bao, *Purdue University*; James Bland, *The University of Toledo*; Kwang Soo Cheong, *Johns Hopkins University*; Amanda Cook, *Bowling Green State University*; Renato Corbetta, *University of Alabama at Birmingham*; Sarah Croco, *University of Maryland*; David E. Cunningham, *University of Maryland*; Seyhan Erden, *Columbia University*; José M. Fernández, *University of Louisville*; Luca Flabbi, *Georgetown University*; Mark A. Gebert, *University of Kentucky*; Kaj Gittings, *Texas Tech University*; Brad Graham, *Grinnell College*; Jonathan Hanson, *University of Michigan*; David Harris, *Benedictine College*; Daniel Henderson, *University of Alabama*; Matthew J. Holian, *San Jose State University*; Todd Idson, *Boston University*; Changkuk Jung, *SUNY Geneseo*; Manfred Keil, *Claremont McKenna College*; Subal C. Kumbhakar, *State University of New York at Binghamton*; Latika Lagalo, *Michigan Technological University*; Matthew Lang, *Xavier University*; Jing Li, *Miami University*; Quan Li, *Texas A&M University*; Drew A. Linzer, *Civiqs*; Steven Livingston, *Middle Tennessee State University*; Aprajit Mahajan, *Stanford University*; Brian McCall, *University of Michigan*; Phillip Mixon, *Troy University*; David Peterson, *Iowa State University*; Leanne C. Powner, *Christopher Newport University*; Zhongjun Qu, *Boston University*; Robi Ragan, *Stetson School of Business and Economics*; Stephen Schmidt, *Union College*; Markus P. A. Schneider, *University of Denver*; Sam Schulhofer-Wohl, *Federal Reserve Bank of Minneapolis*; Christina Suthammanont, *Texas A&M University, San Antonio*; Kerry Tan, *Loyola University Maryland*; Robert Turner, *Colgate University*; Martijn van Hasselt, *University of North Carolina—Greensboro*; David Vera, *California State University*; Christopher Way, *Cornell University*;

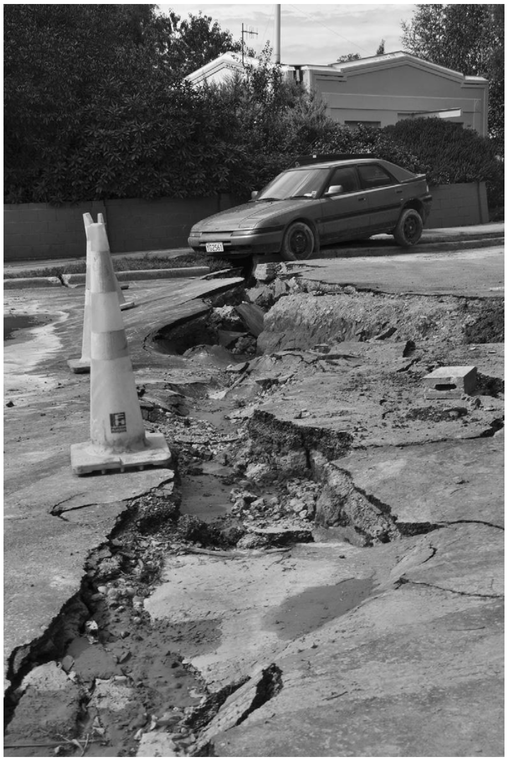
Phanindra V. Wunnava, *Middlebury College*; and Jie Jennifer Zhang, *University of Texas at Arlington*.

I also appreciate the generosity of colleagues who shared data, including Bill Clark, Anna Harvey, Dan Hopkins, and Hans Noel.

The editing team at Oxford has done wonders for this book. Valerie Ashton brought energy and wisdom to the early life of the book. Ann West and Jennifer Carpenter have been supportive and insightful throughout, going to great lengths to make this book the best it can be. Thom Holmes and Maegan Sherlock provided expert development oversight on the first edition. Micheline Frederick has been a very capable editor; without her help this book would have a lot more mistakes . . . and a bit more cussing. Steve Rigolosi and Wesley Morrison performed the unenviable work of copyediting and proofreading my writing with verve. Allison Ball helped with the photo choices, and Tony Mathias has been enthusiastically conveying the message of this book to the marketplace.

I am grateful for the support of my family, Mari, Jack, Emi, and Ken. After years of working on *Real Econometrics*, now we can work on a *real vacation*.

1 The Quest for Causality



How do we know what we know? Or at least, why do we think what we think? The modern answer is evidence. In order to convince others—in order to convince *ourselves*—we need to provide information that others can verify. Something that is a hunch or something that we simply "know" may be important, but it is not the kind of evidence that drives the modern scientific process.

What is the basis of our evidence? In some cases, we can see cause and effect. We see a burning candle tip over and start a fire. Now we know what caused the fire. This is perfectly good knowledge. Sometimes in politics and policy we trace back a chain of causality in a similar way. This process can get complicated, though. Why do some economies stagnate while others thrive? What are the economic and social effects of international trade? Why did Donald Trump win the presidential election in 2016? Why has crime gone down in the United States? For these types of questions, we are not looking only at a single candle; there are lightning strikes, faulty wires, arsonists, and who knows what else to worry about. Clearly, it will be much harder to trace cause and effect.

When there is no way of directly observing cause and effect, we naturally turn to data. And data holds great promise. A building collapses during an earthquake. What about the building led it—and not others in the same city—to collapse? Was it the building material? The height? The design? Age? Location near a fault? While we might not be able to see the cause directly, we can gather information on buildings that did and did not collapse. If the older buildings were more likely

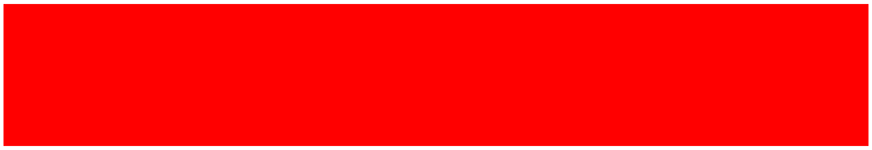


FIGURE 1.1: Rule #1

to collapse, we might reasonably suspect that building age mattered. If buildings constructed without steel reinforcement collapsed no matter what their age, we might reasonably suspect that buildings without reinforcement designs were more likely to collapse.

And yet, we should not get overconfident. Even if old buildings were more likely to collapse, we do not know for certain that age of the building is the main explanation for the collapse. It could be that more buildings from a certain era were designed a certain way; it could be that there were more old buildings in a neighborhood where the seismic activity was most severe. Or the collapse of many buildings that happened to be old could represent a massive coincidence. In other words, correlation is not the same as causation. We put this fact in big blue letters in Figure 1.1 because it is a fundamental starting point in any serious data analysis.

The econometrics we learn in this book will help us to identify causes and make claims about what really mattered—and what didn't. If correlation is not causation, what *does* imply causation? It will take the whole book to fully flesh out the answer, but here's the short version: *if we can find exogenous variation*, then correlation is probably causation. Our task then will be to figure out what exogenous variation means and how to distinguish randomness from causality as best we can.

In this chapter, we introduce three concepts at the heart of the book. Section 1.1 explains the core model we use throughout. Section 1.2 introduces two major challenges that can make it hard to use data to learn about the world. Neither is math. (Really!) The first is randomness: sometimes the luck of the draw will lead us to observe relationships that aren't real; other times random chance will lead us to miss relationships that are real. The second is endogeneity, a phenomenon that can cause us to wrongly think a variable causes some effect when it doesn't. Section 1.3 presents randomized experiments as the ideal way to overcome endogeneity. Usually, these experiments aren't possible, and even when they are, things can go wrong. Hence, the rest of the book is about developing a toolkit that helps us meet (or approximate) the idealized standard of randomized experiments.

1.1 The Core Model

When we talk about cause and effect, we'll refer to the outcome of interest as the **dependent variable**. We'll refer to a possible cause as an **independent variable**.

dependent variable The outcome of interest, usually denoted as *Y*.

independent variable A variable that possibly influences the value of the dependent variable.

The dependent variable, usually denoted as Y, is called that because its value *depends* on the independent variable. The independent variable, usually denoted by X, is called that because it does whatever the hell it wants. It is potentially the cause of some change in the dependent variable.

At root, social scientific theories posit that a change in one thing (the independent variable) will lead to a change in another (the dependent variable). We'll formalize this relationship in a bit, but let's start with an example. Suppose we're interested in the U.S. obesity epidemic and want to analyze the influence of snack food on health. We may wonder, for example, if donuts cause health problems. Our model is that eating donuts (variable X, our independent variable) causes some change in weight (variable Y, our dependent variable). If we can find data on how many donuts people ate and how much they weighed, we might be on the verge of a scientific breakthrough.

Let's conjure up a small midwestern town and do a little research. Figure 1.2 plots donuts eaten and weights for 13 individuals from a randomly chosen town: Springfield, U.S.A. Our raw data is displayed in Table 1.1. Each person has a line in the table. Homer is observation 1. Since he ate 14 donuts per week, $Donuts_1 = 14$. We'll often refer to X_i or Y_i, which are the values of X and Y for person i in the data set. The weight of the seventh person in the data set, Smithers, is 160 pounds, meaning $Weight_7 = 160$, and so forth.

scatterplot A plot of data in which each observation is located at the coordinates defined by the independent and dependent variables.

Figure 1.2 is a **scatterplot** of data, with each observation located at the coordinates defined by the independent and dependent variables. The value of donuts per week is on the X-axis, and weights are on the Y-axis. Just by looking at this plot, we sense there is a positive relationship between donuts and weight because the more donuts eaten, the higher the weight tends to be.

TABLE 1.1 Donut Consumption and Weight

Observation number	Name	Donuts per week	Weight (pounds)
1	Homer	14	275
2	Marge	0	141
3	Lisa	0	70
4	Bart	5	75
5	Comic Book Guy	20	310
6	Mr. Burns	0.75	80
7	Smithers	0.25	160
8	Chief Wiggum	16	263
9	Principal Skinner	3	205
10	Rev. Lovejoy	2	185
11	Ned Flanders	0.8	170
12	Patty	5	155
13	Selma	4	145

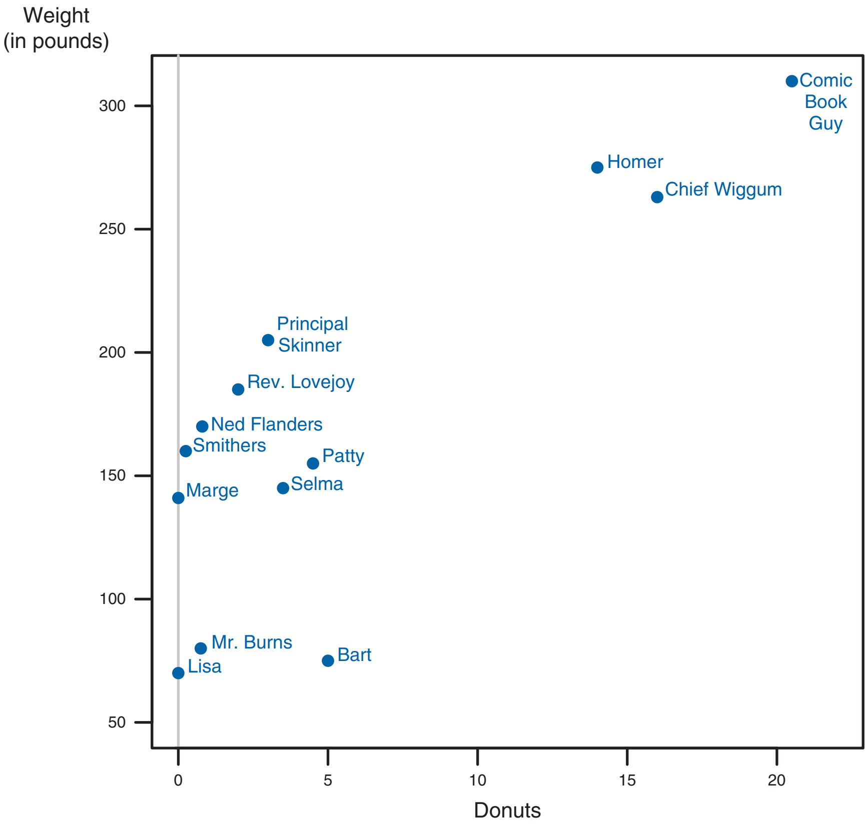


FIGURE 1.2: Weight and Donuts in Springfield

We use a simple equation to characterize the relationship between the two variables:

$$Weight_i = \beta_0 + \beta_1 Donuts_i + \epsilon_i \tag{1.1}$$

- The dependent variable, $Weight_i$, is the weight of person i.

- The independent variable, $Donuts_i$, is how many donuts person i eats per week.

- β_1 is the **slope coefficient** on donuts, indicating how much more[1] a person weighs for each donut eaten. (For those whose Greek is a bit rusty, β is the Greek letter beta.)

- β_0 is the **constant** or **intercept**, indicating the expected weight of people who eat zero donuts.

slope coefficient The coefficient on an independent variable. It reflects how much the dependent variable increases when the independent variable increases by one.

constant The parameter β_0 in a regression model. It is the point at which a regression line crosses the *Y*-axis. Also referred to as the *intercept*.

[1] Or less—be optimistic!

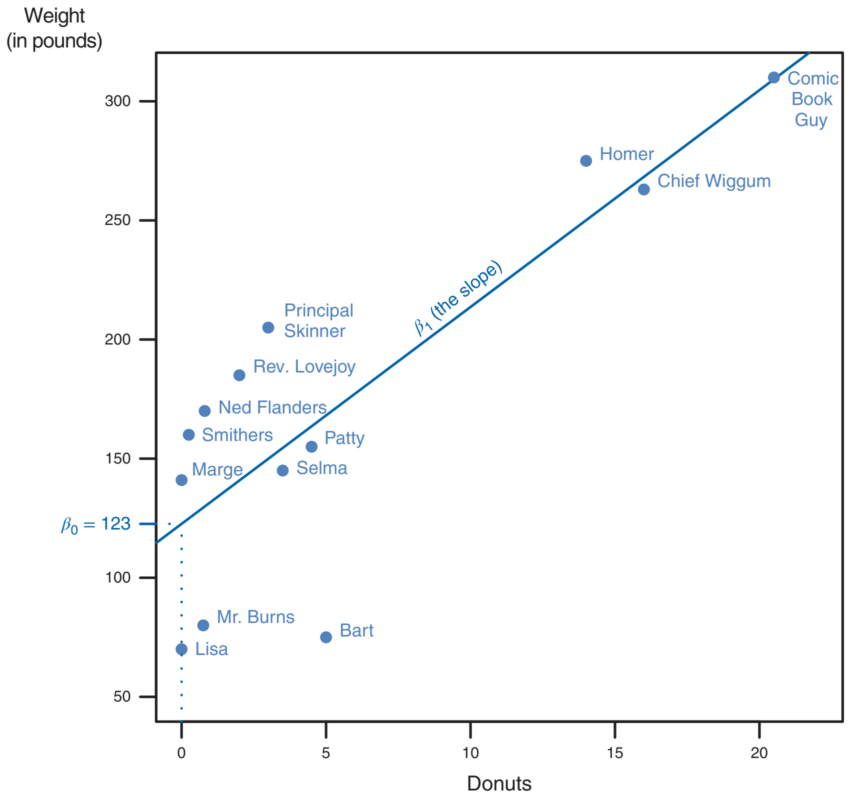


FIGURE 1.3: Regression Line for Weight and Donuts in Springfield

error term The term associated with unmeasured factors in a regression model, typically denoted as ϵ.

- ϵ_i is the **error term** that captures anything else that affects weight. (ϵ is the Greek letter epsilon)

This equation will help us estimate the two parameters necessary to characterize a line. Remember $Y = mX + b$ from junior high? This is the equation for a line where Y is the value of the line on the vertical axis, X is the value on the horizontal axis, m is the slope, and b is the intercept, or the value of Y when X is zero. Equation 1.1 is essentially the same, only we refer to the "b" term as β_0 and call the "m" term β_1.

Figure 1.3 shows an example of a possible line from this model for our Springfield data. The intercept (β_0) is the value of weight when donut consumption is zero ($X = 0$). The slope (β_1) is the amount that weight increases for each donut eaten. In this case, the intercept is about 123, which means that the expected weight for those who eat zero donuts is around 123 pounds. The slope is around 9.1, which means that for each donut eaten per week, weight is about 9.1 pounds higher.

More generally, our core model can be written as

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{1.2}$$

where β_0 is the intercept that indicates the value of Y when $X = 0$ and β_1 is the slope that indicates how much change in Y is expected if X increases by one unit. We almost always care a lot about β_1, which characterizes the relationship between X and Y. We usually don't care a whole lot about β_0. It plays an important role in helping us get the line in the right place, but determining the value of Y when X is zero is seldom our core research interest.

In Figure 1.3, we see that the actual observations do not fall neatly on the line that we're using to characterize the relationship between donuts and weight. The implication is that our model does not perfectly explain the data. Of course it doesn't! Springfield residents are much too complicated for donuts to explain them completely (except, apparently, Comic Book Guy).

The error term, ϵ_i, comes to the rescue by giving us some wiggle room. The error term is what is left over after the variables have done their work in explaining variation in the dependent variable. In doing this service, it plays an incredibly important role for the entire econometric enterprise. As this book proceeds, we will keep coming back to the importance of getting to know our error term.

The error term, ϵ_i, is not simply a Greek letter. It is something real. What it covers depends on the model. In our simple model—in which weight is a function only of how many donuts a person eats—oodles of factors are contained in the error term. Basically, anything else that affects weight will be in the error term: sex, height, other eating habits, exercise patterns, genetics, and on and on. The error term includes everything we haven't measured in our model.

We'll often see ϵ_i referred to as *random error*, but be careful about that one. Yes, for the purposes of the model we are treating the error term as something random, but it is not random in the sense of a roll of the dice. It is random more in the sense that we don't know what the value of it is for any individual observation. But as a practical matter every error term reflects, at least in part, some relationship to real things that we have not measured or included in the model. We will come back to this point often.

REMEMBER THIS

Our core statistical model is

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

1. β_1, the slope, indicates how much change in Y (the dependent variable) is expected if X (the independent variable) increases by one unit.
2. β_0, the intercept, indicates where the regression line crosses the Y-axis. It is the value of Y when X is zero.
3. β_1 is usually more interesting than β_0 because β_1 characterizes a relationship between X and Y.

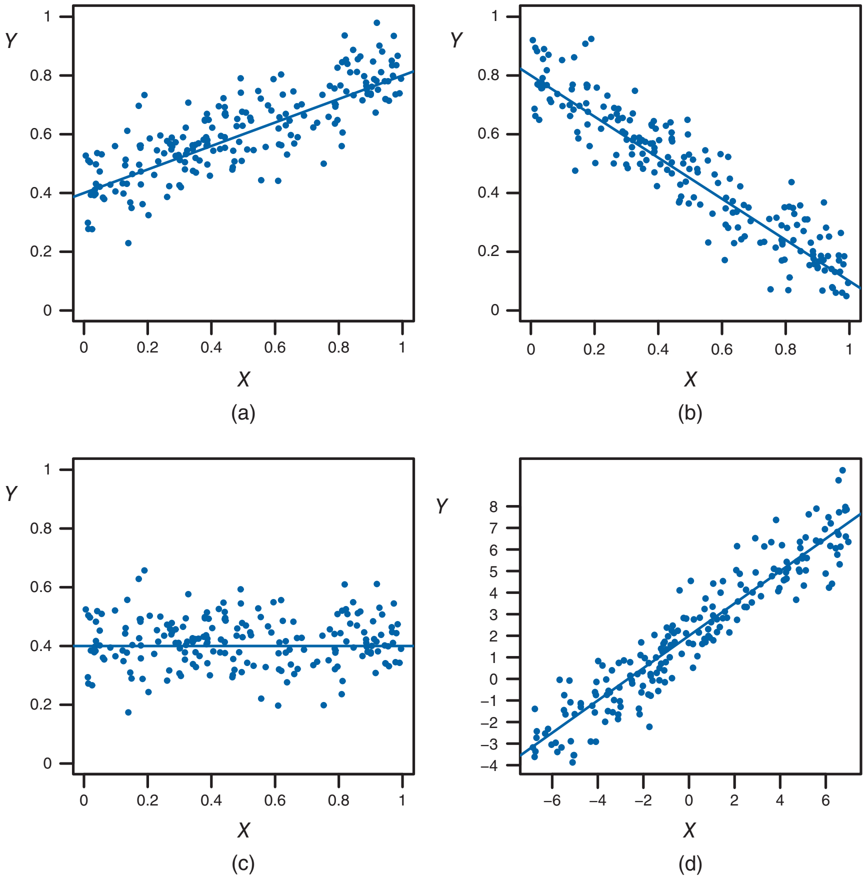


FIGURE 1.4: Examples of Lines Generated by Core Statistical Model (for Review Question)

Review Question

For each of the panels in Figure 1.4, determine whether β_0 and β_1 are greater than, equal to, or less than zero. [Be careful with β_0 in panel (d)!]

1.2 Two Major Challenges: Randomness and Endogeneity

Understanding that there are real factors in the error term helps us be smart about making causal claims. Our data seems to suggest that the more donuts people ate, the more they packed on the pounds. It's not crazy to think that donuts cause weight gain.

But can we be certain that donuts, and not some other factor, cause weight gain? Two core challenges in econometric analysis should make us cautious. One is randomness. Any time we observe a relationship in data, we need to keep in mind that some coincidence could explain it. Perhaps we happened to pick some unusual people for our data set. Or perhaps we picked perfectly representative people, but they happened to have had unusual measurements on the day we examined them.

In the donut example, the possibility of such randomness should worry us, at least a little. Perhaps the people in Figure 1.3 are a bit odd. Perhaps if we had more people, we might get more heavy folks who don't eat donuts and skinny people who scarf them down. Adding those folks to the data set would change the figure and our conclusions. Or perhaps even with the set of folks we observed, we might have gotten some of them on a bad (or a good) day, whereas if we had looked at them another day, we might have observed a different relationship.

Every legitimate econometric analysis therefore will account for randomness in an effort to distinguish results that could happen by chance from those that would be unlikely to happen by chance. The bad news is that we will never escape the possibility that the results we observe are due to randomness rather than a causal effect. The good news, though, is that we can often do a pretty good job characterizing our confidence that the results are not simply due to randomness.

Another major challenge arises from the possibility that an observed relationship between X and Y is actually due to another variable, which causes Y and is associated with X. In the donuts example, worry about scenarios in which we wrongly attribute to our key independent variable (in this case, donut consumption) changes in weight that were caused by other factors. What if tall people eat more donuts? Height is in the error term as a contributing factor to weight, and if tall people eat more donuts, we may wrongly attribute to donuts the effect of height.

There are loads of other possibilities. What if men eat more donuts? What if exercise addicts don't eat donuts? What if people who eat donuts are also more likely to down a tub of Ben and Jerry's ice cream every night? What if thin people can't get donuts down their throats? Being male, exercising, bingeing on ice cream, having itty-bitty throats—all these things are probably in the error term (meaning they affect weight), and all could be correlated with donut eating.

endogenous An independent variable is endogenous if changes in it are related to factors in the error term.

Speaking econometrically, we highlight this major statistical challenge by saying that the donut variable is **endogenous**. An independent variable is endogenous if changes in it are related to factors in the error term. The prefix "endo" refers to something internal, and endogenous independent variables are "in the model" in the sense that they are related to other things that also determine Y (but are not already accounted for by X).

In the donuts example, donut consumption is likely endogenous because how many donuts a person eats is not independent of other factors that influence weight gain. Factors that cause weight gain (e.g., eating Ben and Jerry's ice cream) might be associated with donut eating; in other words, factors that influence the dependent variable Y might also be associated with the independent variable X, muddying the connection between correlation and causation. If we can't be sure that our variation in X is not associated with factors that influence Y, we need to worry about wrongly attributing to X the causal effect of some other variable. We might wrongly conclude that donuts cause weight gain when really donut eaters are more likely to eat tubs of Ben and Jerry's, with the ice cream being the real culprit.

In all these examples, something in the error term that really causes weight gain is related to donut consumption. When this connection exists, we risk spuriously attributing to donut consumption the causal effect of some other factor. Remember, anything not measured in the model is in the error term, and here, at least, we have a wildly simple model in which only donut consumption is measured. So Ben and Jerry's, genetics, and everything else are in the error term.

Endogeneity is everywhere; it's endemic. Suppose we want to know if raising teacher salaries increases test scores. It's an important and timely question. Answering it may seem easy enough: we could simply see if test scores (a dependent variable) are higher in places where teacher salaries (an independent variable) are higher. It's not that easy, though, is it? Endogeneity lurks. Test scores might be determined by unmeasured factors that also affect teacher salaries. Maybe school districts with lots of really poor families don't have very good test scores and don't have enough money to pay teachers high salaries. Or perhaps the relationship is the opposite—poor school districts get extra federal funds to pay teachers more. Either way, teacher salaries are endogenous because their levels depend in part on factors in the error term (like family income) that affect educational outcomes. Simply looking at the relationship of test scores to teacher salaries risks confusing the effect of family income and teacher salaries.[2]

◗ **exogenous** An independent variable is exogenous if changes in it are unrelated to factors in the error term.

The opposite of endogeneity is exogeneity. An independent variable is **exogenous** if changes in it are *not* related to factors in the error term. The prefix "exo" refers to something external, and exogenous independent variables are "outside the model" in the sense that their values are unrelated to other things that also determine Y. For example, if we use an experiment to randomly set the value of X, then changes in X are not associated with factors that also determine Y. This gives us a clean view of the relationship between X and Y, unmuddied by associations between X and other factors that affect Y.

One of our central challenges is to avoid endogeneity and thereby achieve exogeneity. If we succeed, we can be more confident that we have moved beyond correlation and closer to understanding if X causes Y—our fundamental goal. This process is not automatic or easy. Often we won't be able to find purely exogenous variation, so we'll have to think through how close we can get. Nonetheless, the bottom line is this: if we can find exogenous variation in X, we will be in a good position to make reasonable inferences about what will happen to variable Y if we change variable X.

◗ **correlation** Measures the extent to which two variables are linearly related to each other.

To formalize these ideas, we'll use the concept of **correlation**, which most people know, at least informally. Two variables are correlated ("co-related") if they move together. A *positive correlation* means that high values of one variable are associated with high values of the other; a *negative correlation* indicates that high values of one variable are associated with low values of the other.

Figure 1.5 shows examples of variables that have positive correlation [panel (a)], no correlation [panel (b)], and negative correlation [panel (c)].

[2] A good idea is to measure these things and put them in the model so that they are no longer in the error term. That's what we do in Chapter 5.

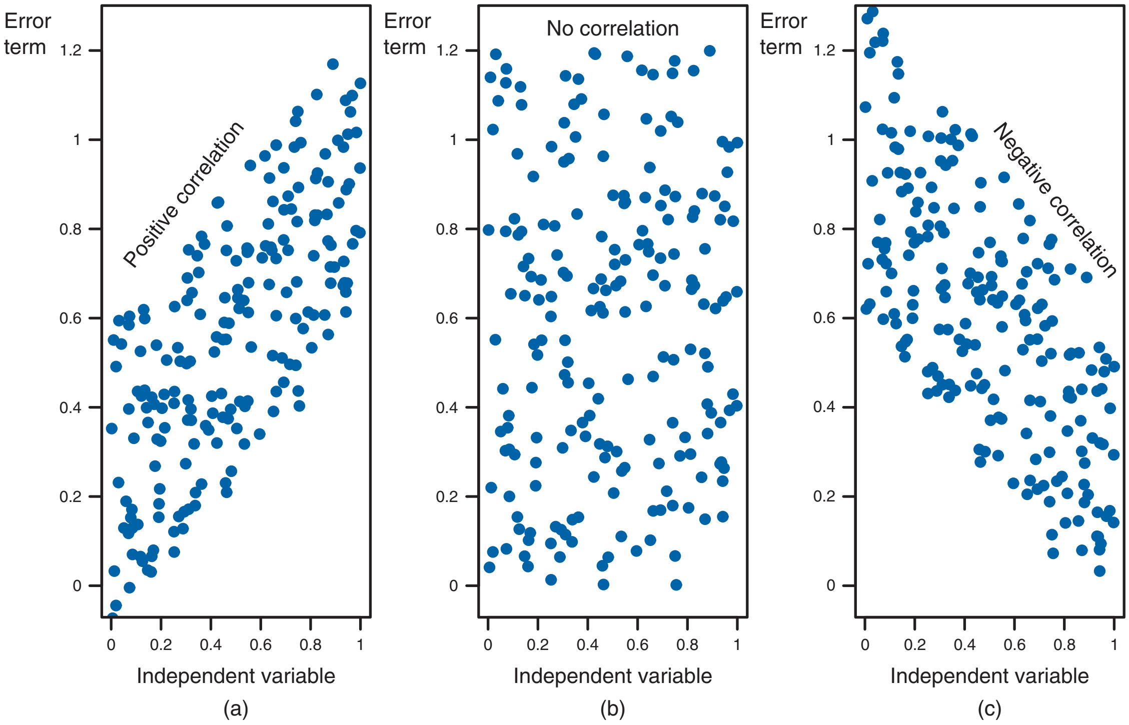


FIGURE 1.5: Correlation

Correlations range from 1 to −1. A correlation of 1 means that the variables move perfectly together.

Correlations close to zero indicate weak relationships between variables. When the correlation is zero, there is no linear relationship between two variables.[3]

We use correlation in our definitions of endogeneity and exogeneity. If our independent variable has a relationship to the error term like the one in panel (a) of Figure 1.5 (which shows positive correlation) or in panel (c) (which shows negative correlation), then we have endogeneity. In other words, we have endogeneity when the unmeasured stuff that constitutes the error term is correlated with our independent variable, and endogeneity will make it difficult to tell whether changes in the dependent variable are caused by our independent variable or the error term.

On the other hand, if our independent variable has no relationship to the error term as in panel (b), we have exogeneity. In this case, if we observe Y rising with X, we can feel confident that X is causing Y.

The challenge is that the true error term is not observable. Hence, much of what we do in econometrics attempts to get around the possibility that something

[3] In Appendix E (page 541), we provide an equation for correlation and discuss how it relates to our ordinary least squares estimates from Chapter 3. Correlation measures linear relationships between variables; we'll discuss non-linear relationships in ordinary least squares on page 221.

unobserved in the error term may be correlated with the independent variable. This quest makes econometrics challenging and interesting.

As a practical matter, we should begin every analysis by assessing endogeneity. First, look away from the model for a moment and list all the things that could determine the dependent variable. Second, ask if anything on the list correlates with the independent variable in the model and explain why it might. That's it. Do that, and we are on our way to identifying endogeneity.

REMEMBER THIS

1. There are two fundamental challenges in econometrics: randomness and endogeneity.
2. Randomness can produce data that suggests X causes Y even when it does not. Randomness can also produce data that suggests X does not cause Y even when it does.
3. An independent variable is endogenous if it is correlated with the error term in the model.
   - **(a)** An independent variable is exogenous if it is not correlated with the error term in the model.
   - **(b)** The error term is not observable, making it a challenge to know whether an independent variable is endogenous or exogenous.
   - **(c)** It is difficult to assess causality for endogenous independent variables.

Discussion Questions

1. Each panel of Figure 1.6 on page 12 shows relationships among three variables: X is an observed independent variable, ϵ is a variable reflecting some unobserved characteristic, and Y is the dependent variable. (In our donut example, X corresponds to the number of donuts eaten, ϵ corresponds to an unobserved characteristic such as exercise, and Y corresponds to the outcome of interest, which is weight.) If an arrow connects X and Y, then X has a causal effect on Y. If an arrow connects ϵ and Y, then the unobserved characteristic has a causal effect on Y. If a double arrow connects X and ϵ, then these two variables are correlated (and we won't worry about which causes which).

   For each panel, explain whether endogeneity will cause problems for an analysis of the relationship between X and Y. For concreteness, assume X is grades in college, ϵ is IQ, and Y is salary at age 26.
2. Come up with your own independent variable, unmeasured error variable, and dependent variable. Decide which of the panels in Figure 1.6 best characterizes the relationship of the variables you chose, and discuss the implications for econometric analysis.

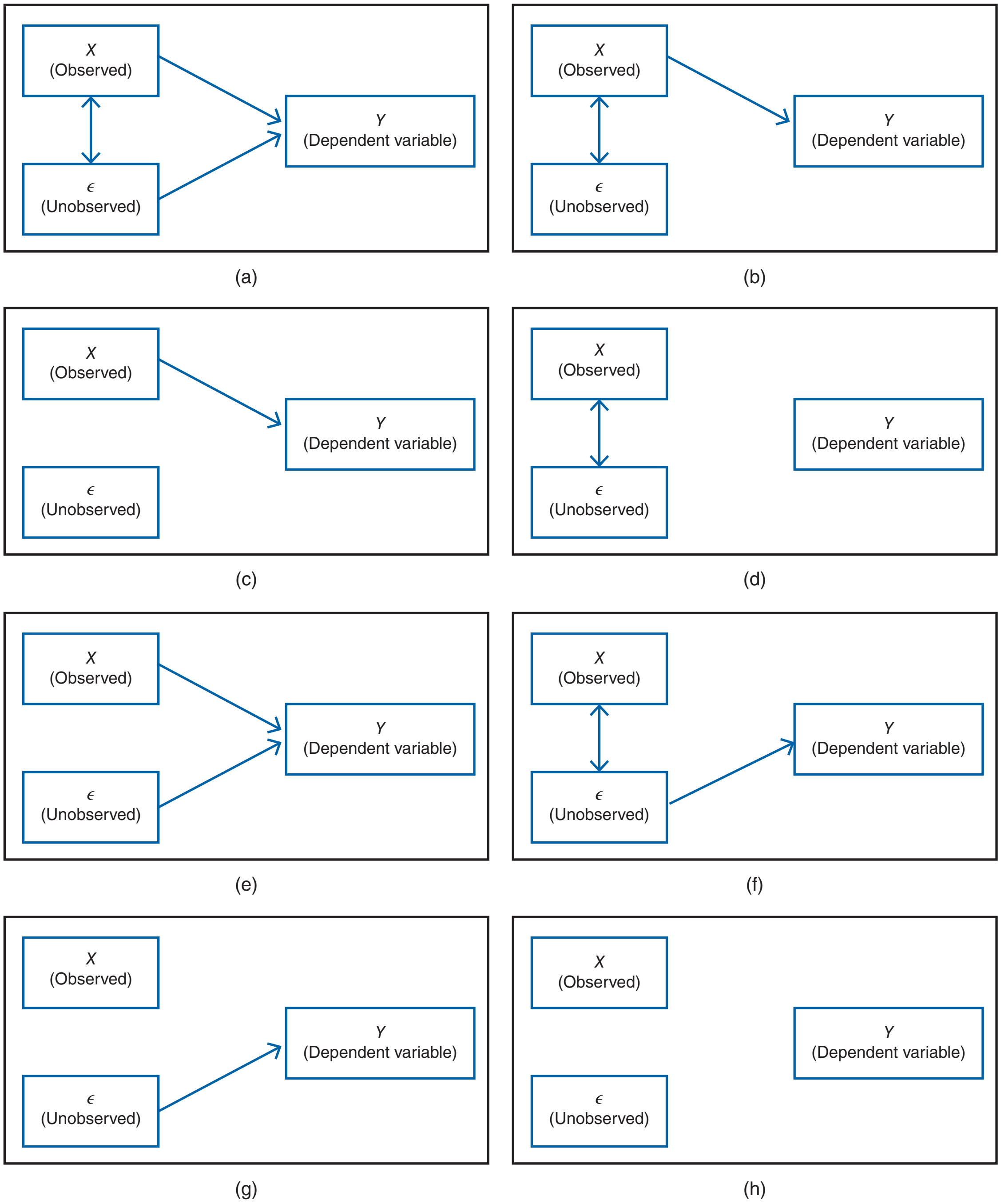


FIGURE 1.6: Possible Relationships Between X, ϵ, and Y (for Discussion Questions)

CASE STUDY Flu Shots

A great way to appreciate the challenges raised by endogeneity is to look at real examples. Here is one we all can relate to: Do flu shots work?

No one likes the flu. It kills about 36,000 people in the United States each year, mostly among the elderly. At the same time, no one enjoys schlepping down to some hospital basement or drugstore lobby, rolling up a shirt sleeve, and getting a flu shot. Nonetheless, every year 100,000,000 Americans dutifully go through this ritual.

The evidence that flu shots prevent people from dying from the flu must be overwhelming, right? Suppose we start by considering a study using data on whether people died (the dependent variable) and whether they got a flu shot (the independent variable):

$$Death_i = \beta_0 + \beta_1 Flu\ shot_i + \epsilon_i \tag{1.3}$$

where $Death_i$ is a (creepy) variable that is 1 if person i died in the time frame of the study and 0 if he or she did not. $Flu\ shot_i$ is 1 if the person i got a flu shot and 0 if not.[4]

A number of studies have done essentially this analysis and found that people who get flu shots are less likely to die. According to some estimates, those who receive flu shots are as much as 50 percent less likely to die. This effect is enormous. Going home with a Band-Aid that has a little bloodstain is worth it after all.

But are we convinced? Is there any chance of endogeneity? If there exists some factor in the error term that affected whether someone died and whether he or she got a flu shot, we would worry about endogeneity.

What is in the error term? Goodness, lots of things affect the probability of dying: age, health status, wealth, cautiousness—the list is immense. All these factors and more are in the error term.

How could these factors cause endogeneity? Let's focus on overall health. Clearly, healthier people die at a lower rate than unhealthy people. If healthy people are also more likely to get flu shots, we might erroneously attribute life-saving power to flu shots when perhaps all that is going on is that people who are healthy in the first place tend to get flu shots.

It's hard, of course, to get measures of health for people, so let's suppose we don't have them. We can, however, speculate on the relationship between health and flu shots. Figure 1.7 shows two possible states of the world. In each figure we plot flu-shot status on the X-axis. A person who did not get a flu shot is in the 0 group; someone who got a flu shot is in the 1 group. On the Y-axis we plot health

[4] We discuss dependent variables that equal only 0 or 1 in Chapter 12 and independent variables that equal 0 or 1 in Chapter 6.

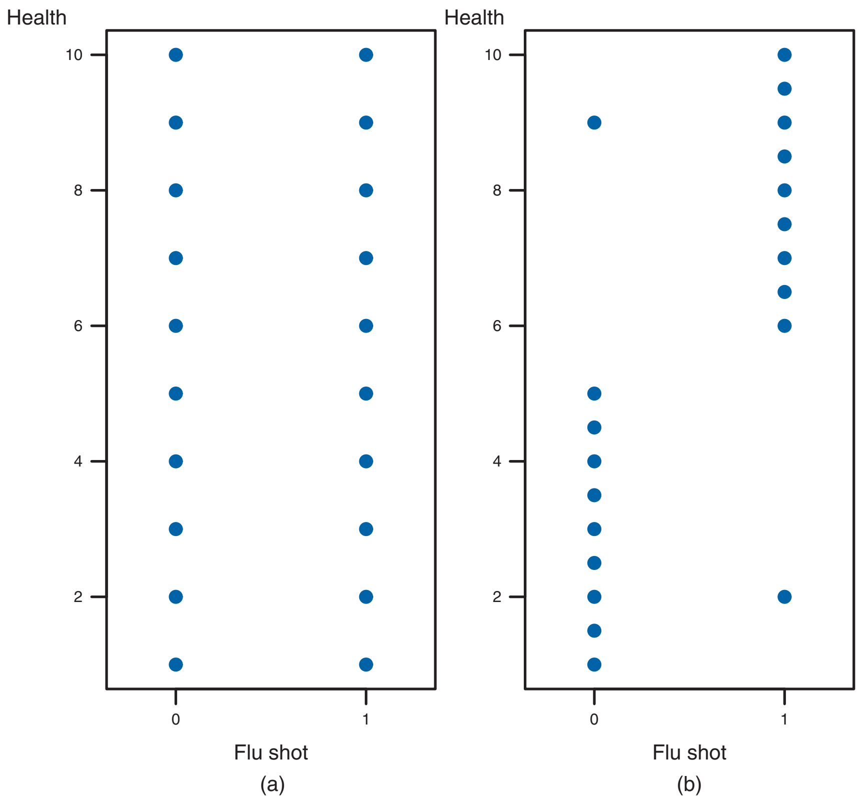


FIGURE 1.7: Two Scenarios for the Relationship between Flu Shots and Health

related to everything but flu (supposing we could get an index that factors in age, heart health, absence of disease, etc.). In panel (a) of Figure 1.7, health and flu shots don't seem to go together; in other words the correlation is zero. If panel (a) represents the state of the world, then our results that flu shots are associated with lower death rates is looking pretty good because flu shots are not reflecting overall health. In panel (b), health and flu shots do seem to go together, with the flu shot population being healthier. In this case, we have correlation of our main variable (flu shots) and something in the error term (health).

Brownlee and Lenzer (2009) discuss some indirect evidence suggesting that flu shots and health are actually correlated. A clever approach to assessing this matter is to look at death rates of people in the summer. The flu rarely kills people in the summer, which means that if people who get flu shots also die at lower rates in the summer, it is because they are healthier overall. And if people who get flu shots die at the same rates as others during the summer, it would be reasonable to suggest that the flu-shot and non-flu-shot populations have similar health. It turns out that people who get flu shots have an approximately 60 percent lower probability of dying outside the flu season.

Other evidence backs up the idea that healthier people get flu shots. As it happened, vaccine production faltered in 2004, and 40 percent fewer people got vaccinated. What happened? Flu deaths did not increase. And in some years, the flu vaccine was designed to attack a set of viruses that turned out to be different from the viruses that actually spread; again, there was no clear change in mortality. This data suggests that people who get flu shots may live longer because getting flu shots is associated with other healthy behavior, such as seeking medical care and eating better.

The point is not to put us off flu shots. We've discussed only mortality—whether people die from the flu—not whether they're more likely to contract the virus or stay home from work because they are sick.[5] The point is to highlight how hard it is to really know if something (in this case, a vaccine) works. If something as widespread and seemingly straightforward as a flu shot is hard to assess definitively, think about the care we must take when trying to analyze policies that affect fewer people and have more complicated effects.

CASE STUDY Country Music and Suicide

Does music affect our behavior? Are we more serious when we listen to classical music? Does bubblegum pop make us bounce through the halls? Both ideas seem plausible, but how can we know for sure?

Stack and Gundlach (1992) looked at data to assess one particular question: Does country music depress us? They argued that country music, with all its lyrics about broken relationships and bad choices, may be so depressing that it increases suicide rates.[6] We can test this claim with the following statistical model:

$$\textit{Suicide rates}_i = \beta_0 + \beta_1 \textit{Country music}_i + \epsilon_i \tag{1.4}$$

where *Suicide rates*$_i$ is the suicide rate in metropolitan area *i* and *Country music*$_i$ is the proportion of radio airtime devoted to country music in metropolitan area *i*.[7]

It turns out that suicides are indeed higher in metropolitan areas where radio stations play more country music. But do we believe this is a *causal* relationship?

[5] Demicheli, Jefferson, Ferroni, Rivetti, and Di Pietrantonj (2018) summarize 52 randomized controlled trials of flu vaccines and conclude that the vaccines reduce the incidence of flu in healthy adults from 2.3 to 0.9 percent. The flu vaccine also reduces the incidence of flu-like illness from 21.5 to 18.1 percent. The effect on hospitalization is not large and not statistically significant. There is no evidence of reducing days off of work. See also DiazGranados, Denis, and Plotkin (2012) as well as Osterholm, Kelley, Sommer, and Belongia (2012).

[6] Really, this is an actual published paper.

[7] Their analysis is based on a more complicated model, but this is the general idea.

(In other words, is country music exogenous?) If radio stations play more country music, should we expect more suicides?

Let's work through this example.

What does β_0 mean? What does β_1 mean? In this model, β_0 is the expected level of suicide in metropolitan areas that play no country music. β_1 is the amount by which suicide rates change for each one-unit increase in the proportion of country music played in a metropolitan area. We don't know what β_1 is; it could be positive (suicides increase), zero (no relation to suicides), or negative (suicides decrease). For the record, we don't know what β_0 is either, but since this variable does not directly characterize the relationship between music and suicides the way β_1 does, we are less interested in it.

What is in the error term? The error term contains factors that are associated with higher suicide rates, such as alcohol and drug use, availability of guns, divorce and poverty rates, lack of sunshine, lack of access to mental health care, and probably many more.

What are the conditions for *X* to be endogenous? An independent variable is endogenous if it is correlated with factors in the error term. Therefore, we need to ask whether the amount of country music played on radio stations in metropolitan areas is correlated with drinking, drug use, and all the other stuff in the error term.

Is the independent variable likely to be endogenous? Are booze, divorce, and guns likely to be correlated to the amount of country music someone has listened to? Have you listened to any country music? Drinking and divorce come up now and again. Could this music appeal more in areas where people drink too much and get divorced more frequently? (To complicate matters, country music could decrease suicide because it lauds family and religion more than many other types of music.) Or could it simply be that people in rural areas who like country music also have a lot of guns? All of these factors—alcohol, divorce, and guns—are plausible influences on suicide rates. To the extent that country music is correlated with any of them, the country music variable would be endogenous.

Explain how endogeneity could lead to incorrect inferences. Suppose for a moment that country music has no effect whatsoever on suicide rates, but that regions with lots of guns and drinking also have more suicides and that people in these regions also listen to more country music. If we look only at the relationship between country music and suicide rates, we will see a positive relationship: places with lots of country music will have higher suicide rates, and places with little country music will have lower suicide rates. The explanation could be that the country music areas have lots of drinking and guns and the areas with little country music have less drinking and fewer guns. Therefore, while it may be correct to say there are more suicides in places where there is more country music, it would be incorrect to conclude that country music *causes* suicides. Or, to put it in another way, it would be incorrect to conclude that we would save lives by banning country music.

As it turns out, Snipes and Maguire (1995) account for the amount of guns and divorce in metropolitan areas and find no relationship between country music and metropolitan suicide rates. So there's no reason to turn off the radio and put away those cowboy boots.

Discussion Questions

1. Labor economists often study the returns on investment in education (see, e.g., Card 1999). Suppose we have data on salaries of a set of people, some of whom went to college and some of whom did not. A simple model linking education to salary is

$$Salary_i = \beta_0 + \beta_1 College\ graduate_i + \epsilon_i$$

where the value of $Salary_i$ is the salary of person i and the value of $College\ graduate_i$ is 1 if person i graduated from college and is 0 if person i did not.

   (a) What does β_0 mean? What does β_1 mean?

   (b) What is in the error term?

   (c) What are the conditions for the independent variable X to be endogenous?

   (d) Is the independent variable likely to be endogenous? Why or why not?

   (e) Explain how endogeneity could lead to incorrect inferences.

2. Donuts aren't the only food that people worry about. Consider the following model based on Solnick and Hemenway (2011):

$$Violence_i = \beta_0 + \beta_1 Soft\ drinks_i + \epsilon_i$$

where $Violence_i$ is the number of physical confrontations student i was in during a school year and $Soft\ drinks_i$ is the average number of cans of soda student i drinks per week.

   (a) What does β_0 mean? What does β_1 mean?

   (b) What is in the error term?

   (c) What are the conditions for the independent variable X to be endogenous?

   (d) Is the independent variable likely to be endogenous? Why or why not?

   (e) Explain how endogeneity could lead to incorrect inferences.

3. We know U.S. political candidates spend an awful lot of time raising money. And we know they use the money to inflict mind-numbing ads on us. Do we know if the money and the ads