Data and the American Dream

Contemporary Social Controversies and the American Community Survey

Matthew J. Holian

palgrave
macmillan

Data and the American Dream

Matthew J. Holian

Data and the American Dream

Contemporary Social Controversies and the American Community Survey

palgrave
macmillan

Matthew J. Holian
Professor of Economics
San Jose State University
San Jose, CA, USA

ISBN 978-3-030-64261-7 ISBN 978-3-030-64262-4 (eBook)
https://doi.org/10.1007/978-3-030-64262-4

This Palgrave Macmillan imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

To Bridget

Preface

The main goal of this book is to show the reader how to use core empirical methods in social science and econometric research. I do this by presenting case studies of published scholarly journal articles, organized into the following areas: housing, migration, labor, health care, family, and transportation. Empirical research can shed badly needed light on many contemporary social controversies, from climate change to illegal immigration to health care. The book concludes by describing how careful empirical estimates can guide decision making, through *cost-benefit analysis*, to find public policies that lead to greater happiness while accounting for environmental, public health, and other impacts.

To illustrate econometric research methods, this book describes empirical studies that have in common the use of the same underlying data: individual responses to the American Community Survey (ACS), the nation's largest household survey. This book was written for student and professional audiences, and self-directed learners. It has a website where I make available replication files in R and Stata format for all of the statistics discussed in the book. Another novel feature of this book is that the replication files all draw from a single master data file, available on this book's website, or from IPUMS, an important data center housed at the University of Minnesota. The design of this book illustrates the multitude of ways to use the ACS data in research, as well as empirical best practices.

This book could be used as a supplemental text in introductory undergraduate or graduate econometrics courses, or as the main text in a course where students also read the original studies this book draws

upon. I discuss sample course structures and suggested textbook pairings in Appendix A, which also contains a guide to the free R software, downloading the ACS and other *public-use microdata*, and running the replication files, which assumes little background knowledge on the part of the reader. The only prerequisite to reading this book is a course in introductory statistics.

I emphasize an intuitive understanding of the statistical techniques used in modern empirical economics, as a complement to the exposition in the leading textbooks. Each of the book's four parts cover distinct econometric concepts, and start with a list of learnings goals. I include a set of review questions at the end of each chapter that reinforce the learning goals. Finally, a glossary containing definitions of key terms, which are noted throughout the text in *italics*, can help the reader make sense of the confusing econometrics jargon one finds in published economics research.

I also try to keep the tone light to make the book accessible to both student and professional audiences. To bring the survey data to life, I include stories about some of the survey's target population–Americans–including some about me and my household. My aim in writing this book was to introduce students to the methods of modern econometric analysis, in a way that lights the fire of interest in beginning students to do their own research, while still being informative and thought-provoking for professionals already working in the field.

The tone is light but knowing how to apply the techniques covered is a valuable skill. Readers will see how research results published in top scholarly journals often use relatively simple techniques, such as the calculation of *means*, and the *difference in means* between two groups. The techniques one most often encounters in econometric studies using the ACS involve the *difference-in-differences* of means, and *regression control*. These are also the focus here. One study I discuss uses a technique called *instrumental variables*. Many applications of the more advanced techniques can get exceedingly complicated, but I have carefully selected examples of methods in their most basic forms. Beginning students are easily distracted by bells and whistles when they first encounter econometrics; this book and its website break these results down to their core, while opening up the hood on the techniques used by top scholars.

The explosion of data from web transactions has generated substantial interest in *big data analytics*, but what is the best way to teach students how to do it? Today, students and researchers can access public use microdata on over 2 billion respondent records from structured surveys from

over 100 countries, dating from 1703 to the present.[1] My view is that microdata like the ACS provides a better introduction to data analysis than does aggregate or unstructured data, because microdata is easy to understand intuitively; we can imagine ourselves as survey respondents. In my teaching, I have adopted the increasingly popular *replicate and extend* approach, and this book is designed to be used in classes that either take this pedagogical approach or that otherwise focus on doing research. I've used it in both graduate and undergraduate courses. The idea behind replicate and extend is, once a student is able to run an analysis file that replicates a study using the raw ACS data, such as the R scripts that are available on this book's webpage, it's not that hard to modify the script in a way that does something original. As students gain skills and confidence, they can start to replicate studies on their own, until one day, they do research that others replicate.

The outline of this book is as follows. Chapter 1 describes the ACS and how to use microdata from it to calculate descriptive statistics and make inferences about cause and effect social relationships. It introduces the core statistical technique of regression. This chapter emphasizes an intuitive understanding of techniques and concepts, and defines and clarifies dozens of key terms used in econometric research. Questions for Review at the end of the chapter, on topics including sample weighting and inflation adjustments, illustrate the use of empirical best practices to those readers either beginning in econometrics or with experience but looking to add a valuable new data source to their repertoire.

Chapter 2 illustrates the regression control technique for causal inference, through an empirical case study of building codes and household energy consumption. It describes and defines key concepts, like logged variables and fixed effects, so that the beginning reader can both understand and use the regression control technique. This chapter also describes the research process, and the path a researcher can take from replicating a study, to extending it and doing original research based on the study. Questions for Review at the end of the chapter walk the reader through this process, from downloading data and replicating a published research study, to modifying the computer code and creating new knowledge.

[1] Ruggles, Steven. "Big microdata for population research." *Demography* 51, no. 1 (2014): 287–297.

Chapter 3 is the first of three chapters on the Difference-in-Differences (D-in-D) technique. Through an empirical case study of the effect of immigration policy on employment among Salvadoran immigrants, it illustrates how natural experiments can be analyzed using the ACS data and the regression model introduced in earlier chapters. This chapter introduces the basic D-in-D model, and a variant of it, the basic D-in-D model with control variables. It also reviews some of the extensive literature that has analyzed both international migration and migration within the United States and its cities, using the ACS data. End of chapter Review Questions reinforce the concepts introduced in the chapter, and give the reader ideas for original research they can carry out.

Chapter 4 is the second chapter on the D-in-D technique. An empirical case study of the Affordable Care Act on entrepreneurship offers another illustration of the basic D-in-D model. This chapter introduces a new way, called pre-trends analysis, to probe the model's assumptions, and a new variant of D-in-D, the fixed effect D-in-D model. It also reviews some of the extensive literature that has analyzed both health and labor market topics using the ACS data. This chapter also revisits a descriptive study on lawyer earnings, first introduced in Chapter 1, and extends it to software developer earnings. End of chapter questions are designed to give the reader ideas for original research on labor and health questions.

Chapter 5 is the third chapter on the D-in-D technique. It also revisits the technique of regression control, first introduced in Chapters 1 and 2, through a discussion of the effect of marriage and children on female labor market earnings. An empirical case study of the Great Recession on fertility offers a creative illustration of both the basic D-in-D model and ways to probe the validity of its assumptions. It also reviews some of the extensive literature that has analyzed family topics using the ACS data.

The brief Chapter 6 illustrates the instrumental variables (IV) technique for causal inference, through an empirical case study of land-use, as measured by population density, and vehicle ownership. It describes how the ACS can be used to study questions related to commuting and working from home. It emphasizes an intuitive understanding of the technique, and shows how a natural experiment on sibling gender and home size, first introduced in Chapter 1, can be used in an IV model.

The concluding Chapter 7 describes how empirical results obtained with the econometric techniques emphasized in preceding chapters can be used to make policy recommendations. It introduces the technique of Cost-Benefit Analysis (CBA) which is a decision making technique that

requires having good empirical estimates. It revisits the topic of building energy codes, first introduced in Chapter 2, to illustrate CBA through a case study. Although no decision making technique is perfect, this chapter argues that empirical researchers should have a greater familiarity with CBA so that their policy recommendations are more likely to rationally account for a comprehensive set of impacts.

This book also has two appendices. Appendix A, as already suggested, provides a link between the more intuitive treatment found in this book and the more formal treatment found in leading textbooks. It also contains a guide to the free R software, downloading the ACS, running the replication files, guidance on which types of studies are good candidates to replicate, and a list of lessons learned regarding best practices in the analysis of the ACS microdata. Appendix B contains the ACS survey instrument, as it appeared in 2015. It can be extremely revealing to see the actual wording of questions, and this will help a reader to understand what information is gathered by the survey.

This book describes work by dozens of economists and other researchers. At various points in writing this book, I emailed some of these authors asking if they would share their code with me. I discuss this aspect of writing the book in more detail in Appendix A. I thank all of the too-many-to-name authors who corresponded with me by email. I am grateful to the authors who could share their code with me. Among them I especially thank John Winters, whose research on earnings by college major inspired the case study in Chapter 1, as well as Table 4.1, James Bailey, whose work on health insurance is featured in the case study in Chapter 4, and Matthew Kotchen, whose economic analysis enabled my conclusion in Chapter 7. The articles for two other case studies were published in the *American Economic Review* (AER) journal, with research data and code in Stata format. I applaud this journal's policy of requiring authors to submit replication files. Not all authors that publish in the AER submit replication files that show how to use the raw data to arrive at the estimates reported in the article. I therefore thank Dora Costa and Matt Kahn, whose work on building energy codes is featured in Chapter 2, and Pia Orrenius and Madeline Zavodny, whose paper on immigration policy is featured in Chapter 3, for producing transparent replication files for the research community.

Many of my students at San Jose State University, through their term papers and associated R scripts, helped me decide which studies to replicate and describe in this book. Austin Tse and Rosalyn Hua deserve special

mention for directly writing some of the R code that appears in the online replication files, but there are many others who helped improve this book. My colleagues at SJSU, especially Darwyyn Deyo, Colleen Haight, and Paul Lombardi, provided valuable feedback on draft chapters. I thank Gordon Douglas for allowing me to present the early stage manuscript in his urban studies working group. Scott Cunningham inspired me to write this book, Nic Albert and Andrew Chang helped me form the idea for it, and my university granted me a sabbatical leave for the Fall 2019 semester, during which time I wrote the bulk of this book.

I would like to thank my editor at Palgrave Macmillan Elizabeth Graber who identified the potential in this project and provided guidance that improved it. Shreenidhi Natarajan and the production team at Palgrave was a pleasure to work with. A portion of Chapter 7 previously appeared in an edited volume published by the Center for Growth and Opportunity titled, *Regulation and Economic Opportunity: Blueprints for Reform* and I thank CGO for allowing me to use it here. Two anonymous reviewers provided many useful comments on this book in the proposal stage, and I deeply appreciate the valuable comments and suggestions from the reviewer who carefully read the final manuscript, which pushed me to improve it. Of course, any errors remain my own.

One of my hopes with this book is that it will be widely used by students and independent learners, who will both learn from and improve upon the coding I have done. This book's web page contains a form readers can use to submit their improvements, and also replications they have carried out of studies that use the ACS or related data. I plan to update the web page with links to some of the replications produced by the community of users of this book.

San Francisco, USA Matthew J. Holian

Contents

Acronyms and Abbreviations

AC	Air conditioner
ACA	Affordable Care Act
ACS	American Community Survey
AER	American Economic Review
BCA	Benefit–Cost Analysis
BD	Bailey and Dave
CAFE	Corporate Average Fuel Economy
CBA	Cost–Benefit Analysis
CBK	Text "Codebook" file describing an IPUMS extract
CO2	Carbon Dioxide
CPS	Current Population Survey
CSV	Comma Separated Value
DACA	Deferred Action for Childhood Arrivals
DDI	Data Documentation Initiative
D-in-D	Difference-in-Differences
EIA	Economic Impact Analysis
FIA	Fiscal Impact Analysis
GB	Gigabyte
GHG	Greenhouse Gas
GIS	Geographic Information Systems
IPCC	Intergovernmental Panel on Climate Change
IPUMS	Integrated Public Use Microdata Series
IV	Instrumental Variables
JK	Jacobsen and Kitchen
kWh	Kilowatt Hour
NOx	Nitrogen Oxide

NPV	Net Present Value
OZ	Orennius and Zavodny
PC	Personal Computer
PDF	Portable Document Format
PM2.5	Particulate Matter
PUMA	Public Use Microdata Area
PUMS	Public Use Microdata Sample
RIA	Regulatory Impact Analysis
SO2	Sulfur Dioxide
STEM	Science, Technology, Engineering, and Mathematics
TPS	Temporary Protected Status
TWFE	Two-Way Fixed Effects

List of Figures

List of Tables

PART I

Descriptive Statistics, Causal Inference, and Regression

Learning Goals for Part I

1. Explain how to use the ACS microdata to calculate a statistic for a population in a geographic area.
2. Compare and contrast descriptive statistics, inferential statistics, and causal inference.
3. Give an example of a difference in means estimated with the ACS that cannot be interpreted as a causal effect, and one that can.
4. Explain how to use a bivariate regression model to estimate a difference in means.
5. Using bivariate and multivariate regression models, evaluate a regression coefficient estimate in terms of omitted variable bias.

CHAPTER 1

Introduction: Stories, Data and Statistics

Each person in America and around the world has a story. Government agencies like the U.S. Census Bureau systematically collect these stories by conducting large-scale surveys, and archive millions per year as data that are publicly-available online. Secondary researchers then use these data on individuals and the households in which they live to calculate statistics. As an example, a researcher might use data to calculate that average household size was 2.44 in the US in 2015, although it is noticeably lower in one Manhattan neighborhood at 1.68 people per household.

Quite often, journalists and others use these statistics as evidence of real or perceived cause and effect relationships. For example, someone may use the statistics on household size and present them as evidence that living in an expensive, high-density area like Manhattan actually causes families to have fewer children. But there are other possible explanations, including that smaller families are more likely to move to Manhattan in the first place. This book describes how to calculate statistics using individual-level population data, and how (and how not) to use the statistics.

This book presents an accessible discussion of economic and social science research that share the theme of utilizing one data source: individual responses to the American Community Survey (ACS), the nation's largest household survey. The people represented in our data sets have real lives and aren't "just statistics" or even just data. But at the same time, the individual response data—also called *microdata*—is incredibly useful for understanding contemporary social life as we can look in the data set and

M. J. Holian, *Data and the American Dream*,
https://doi.org/10.1007/978-3-030-64262-4_1

find examples of, say, a Spanish-speaking single-mom with three daughters, and learn many details about their lives. This book presents evidence for concerned citizens on important social issues like climate change, health care, and immigration, while highlighting best practices in social science research for students and practitioners.

The stories in this book draw from many sources including the experience of my own family, families we have met living in San Francisco, and the lives of students I have met teaching for the last dozen years at San Jose State University. I grew up in a middle-class Ohio suburb and as a result many of my own experiences are typical, which makes them relevant in a book about American social life. I also happen to now have a unique vantage point from which to view the world's most dynamic economic region, as a professor at the oldest public university in California, which today is in the heart of Silicon Valley.

This introductory chapter describes the plan of the book, important background information on the ACS, and illustrates key statistical techniques, and is followed by five chapters with topical themes: homes, migration, work, family, transportation. A typical topics chapter presents a detailed case of one study that uses the ACS, and all chapters draw from other studies that use the ACS, other scholarly sources, the news media, and popular culture. A concluding chapter shows how credible causal estimates can be used to make decisions by introducing a framework called Cost–Benefit Analysis (CBA). Appendix A is an important part of this book. It describes software, data and online resources, and parts will be of interest to both student and professional audiences.

Every 10 years since 1790, the U.S. government has taken a Census. Along with enumerating the population, the U.S. Census Bureau has in recent years taken advantage of the opportunity of counting everyone to ask a subset of Americans questions about their lives, including the occupation of workers in the household, their age, race, and so on. The decennial Census asked these detailed question to about 5% of the population during census years, the so-called "long-form" sample. The data available from the U.S. Census Bureau is one of the oldest and richest anywhere, and anyone with a computer and Internet connection can download data on millions of Americans.

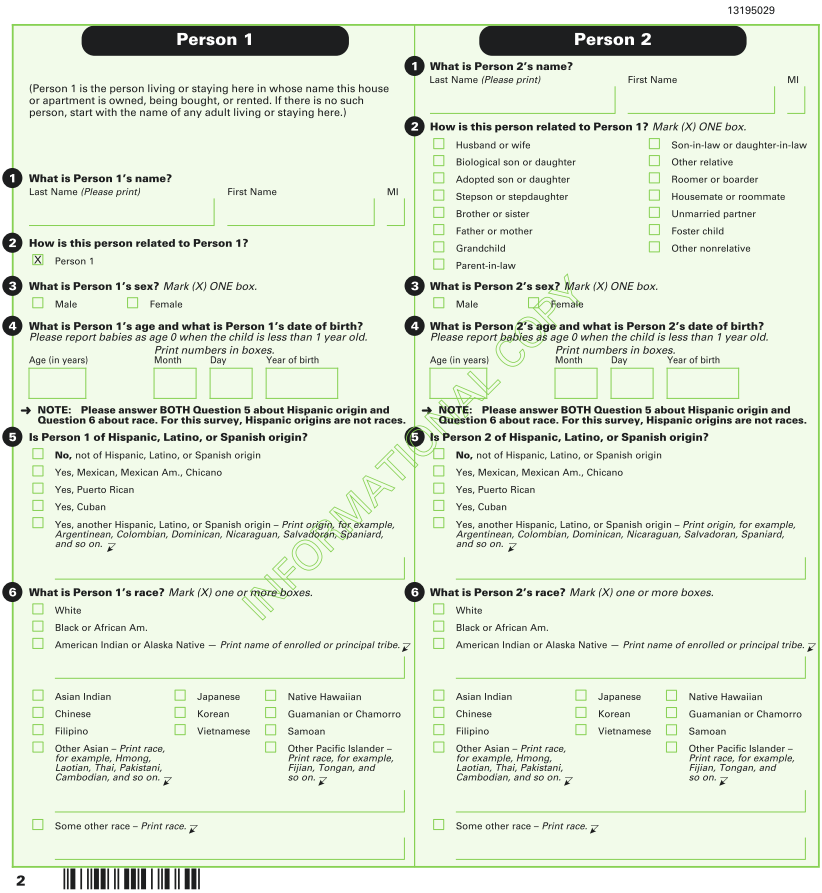
While it is sometimes possible to obtain these data directly from the Census Bureau's webpage, I discuss in Appendix A why it is easiest to download Census microdata from a third-party, rather than from the Census Bureau directly, and in particular to download the data from an organization at the

University of Minnesota that goes by the acronym IPUMS, which stands for, Integrated Public Use Microdata Series. IPUMS distributes Census data going back to 1850 with a user-friendly web page (though the records from 1890 are unavailable as they were lost in a 1921 fire in the Commerce Department Building where the paper records of the time were stored).

The 2000 decennial Census was the last one to incorporate a long-form survey. Since then, the American Community Survey has replaced the decennial Census long form for the purposes of asking respondents about their lives. The Census Bureau still enumerates the population every ten years through a decennial Census, but since 2010 the decennial Census no longer contains a long-form. The beauty of the ACS is that it was modeled on the long-form but samples 1% of households every year, not 5% every decade. The questions on the ACS are still very similar to those that were asked on the 2000 long form, but the questionnaire has evolved some. For example, the ACS now asks college degree holders what they majored in, and all persons their health insurance status. This book focuses on studies of the contemporary period that have used ACS data from years 2004 to 2017, but the methods these studies use are highly applicable to older Census data as well.

To get an idea of what the ACS looks like, consider the information presented in Fig. 1.1. This is a snapshot of the survey form a respondent to the ACS may see. Respondents don't always see these questions, for example, in cases where the survey was conducted through an in-person interview with a Census worker, but this is the form someone completing the survey by mail would see. I include this image so that a reader can imagine they themselves are filling out this survey. The ACS is a long survey and the full survey questionnaire (which is reproduced in Appendix B) is about 15 pages, but seeing the questions just as respondents see them can help a reader imagine how their own stories and experiences can be turned into data. For example, what's your age, gender, and highest level of educational attainment? Some questions are more sensitive: What was your total income from all sources last year? Are you a citizen of the USA?

It may seem surprising that every year, the Census Bureau is able to conduct these surveys with millions of Americans. The ACS aims to survey 1% of the population each year. Given the U.S. population is currently 327 million, this means about 3.27 million people are surveyed every year. Take a moment to reread that line. The ACS is truly a massive survey effort. What is especially remarkable is the very high response rate, which is reliably


13195029

Person 1

(Person 1 is the person living or staying here in whose name this house or apartment is owned, being bought, or rented. If there is no such person, start with the name of any adult living or staying here.)

1 What is Person 1's name?
Last Name (Please print) First Name MI

2 How is this person related to Person 1?
[X] Person 1

3 What is Person 1's sex? Mark (X) ONE box.
Male Female

4 What is Person 1's age and what is Person 1's date of birth?
Please report babies as age 0 when the child is less than 1 year old.
Print numbers in boxes.
Age (in years) Month Day Year of birth

→ NOTE: Please answer BOTH Question 5 about Hispanic origin and Question 6 about race. For this survey, Hispanic origins are not races.

5 Is Person 1 of Hispanic, Latino, or Spanish origin?
No, not of Hispanic, Latino, or Spanish origin
Yes, Mexican, Mexican Am., Chicano
Yes, Puerto Rican
Yes, Cuban
Yes, another Hispanic, Latino, or Spanish origin – Print origin, for example, Argentinean, Colombian, Dominican, Nicaraguan, Salvadoran, Spaniard, and so on.

6 What is Person 1's race? Mark (X) one or more boxes.
White
Black or African Am.
American Indian or Alaska Native — Print name of enrolled or principal tribe.
Asian Indian Japanese Native Hawaiian
Chinese Korean Guamanian or Chamorro
Filipino Vietnamese Samoan
Other Asian – Print race, for example, Hmong, Laotian, Thai, Pakistani, Cambodian, and so on.
Other Pacific Islander – Print race, for example, Fijian, Tongan, and so on.
Some other race – Print race.

Person 2

1 What is Person 2's name?
Last Name (Please print) First Name MI

2 How is this person related to Person 1? Mark (X) ONE box.
Husband or wife
Biological son or daughter
Adopted son or daughter
Stepson or stepdaughter
Brother or sister
Father or mother
Grandchild
Parent-in-law
Son-in-law or daughter-in-law
Other relative
Roomer or boarder
Housemate or roommate
Unmarried partner
Foster child
Other nonrelative

3 What is Person 2's sex? Mark (X) ONE box.
Male Female

4 What is Person 2's age and what is Person 2's date of birth?
Please report babies as age 0 when the child is less than 1 year old.
Print numbers in boxes.
Age (in years) Month Day Year of birth

→ NOTE: Please answer BOTH Question 5 about Hispanic origin and Question 6 about race. For this survey, Hispanic origins are not races.

5 Is Person 2 of Hispanic, Latino, or Spanish origin?
No, not of Hispanic, Latino, or Spanish origin
Yes, Mexican, Mexican Am., Chicano
Yes, Puerto Rican
Yes, Cuban
Yes, another Hispanic, Latino, or Spanish origin – Print origin, for example, Argentinean, Colombian, Dominican, Nicaraguan, Salvadoran, Spaniard, and so on.

6 What is Person 2's race? Mark (X) one or more boxes.
White
Black or African Am.
American Indian or Alaska Native — Print name of enrolled or principal tribe.
Asian Indian Japanese Native Hawaiian
Chinese Korean Guamanian or Chamorro
Filipino Vietnamese Samoan
Other Asian – Print race, for example, Hmong, Laotian, Thai, Pakistani, Cambodian, and so on.
Other Pacific Islander – Print race, for example, Fijian, Tongan, and so on.
Some other race – Print race.

2

Fig. 1.1 The first page of the ACS survey form

more than 90% each year. While many survey organizations are content with much lower response rates, the Census Bureau is able to achieve such high rates, in part because respondents are legally obligated to complete the survey (though I have found no evidence someone has ever been prosecuted for failing to complete the ACS questionnaire). Of course whether everyone is giving completely truthful information is another question, which I address later in this chapter, and not all respondents answer all questions;

Table 1.1 ACS Raw Person Data, 2015, PUMA 068511

SERIAL	*AGE*	*CITIZEN*	*RELATED*	*EDUCD*	*SEX*	*INCTOT*	*UHRSWORK*	*OCC1990*
67205	57	0	101	71	1	72000	40	217
67205	56	0	201	101	2	21000	28	156
67359	52	0	101	81	1	130000	56	417
67359	51	0	201	65	2	60000	40	337
67383	61	3	101	40	1	19000	30	276
67481	40	0	101	101	1	60000	40	103
67481	7	0	301	14	1	9999999	0	999
67481	27	0	1115	101	2	48000	40	379
68781	39	3	101	116	1	115300	40	55
68781	39	3	201	101	2	0	0	999
68781	3	0	301	2	2	9999999	0	999
68781	2	0	301	1	1	9999999	0	999

Notes Data is 2015 ACS microdata for twelve individuals in PUMA 11 in Santa Clara County, California. The Data section of the file `script1.R` on the book's webpage produces a table with these data in it

Appendix A contains a discussion of this issue, which the Census Bureau refers to as "item nonresponse."

After the Census completes a survey, it is digitized. Table 1.1 presents data from twelve people surveyed in one California neighborhood in 2015. There is nothing special about the neighborhood I chose to take these data from, except that it happens to be near my university in San Jose. After completing a fifteen-page survey, there are many things we know about the households who were surveyed, and the questions are turned into dozens of variables. Table 1.1 shows data on just nine variables. A *variable* is a column of data, and the variable name is at the top of the column.

Each row of this table contains information on one of the twelve people. These twelve people live in five separate households, which we know because each household is given a unique identifying value of the variable SERIAL, which stands for "serial number." In this book, I always write variable names in all capital letters, and I have also retained IPUMS variable names, even in cases like with SERIAL where a name like "HOUSEHOLD ID" would have been more descriptive. Table A.4 in the Appendix describes the coding for all of the variables in Table 1.1. At this point, a reader could just read my descriptions of the variables in the text that follows, though later I'll discuss why it would be better to cross-reference the data in Table 1.1 with the *codebook* details in Table A.4.

Consider the first two rows. Both individuals live in the same household 67205, are aged 57 and 56, respectively, and both are U.S. citizens. We

know this last fact about their citizenship because according to the codebook, when the value of the variable CITIZEN equals 0 it signifies the individual was born in the USA and is thus a citizen.

The SEX variable tells us the first individual in Table 1.1 is male, because it takes on a value of 1, and the second is female because it's 2. We know from the relationship variable RELATED they are married; a value of RELATED equal to 101 indicates this is the "reference person." This is the person the Census Bureau interviewed; the reference person is sometimes referred to as the "head of household" but in modern times this designation has less meaning. We see at the top of Fig. 1.1 that the reference person is also referred to as Person 1 and, "...is the person living or staying here in whose name this house or apartment is owned, being bought, or rented." A value of RELATED equal to 201 indicates this person is the spouse of the reference person. They both work; she works 28 hours per week on average and he works a 40-hour week (these are the values of UHRSWORK, which stands for "usual hours work"). His occupation is coded as 217, and when the variable OCC1990 takes this value it indicates the occupation is "Drafters." She works in occupation 156, "Primary school teachers." She is more highly educated; a value of EDUCD equal to 71 indicates he has one year of college but no degree, and her value of this variable is 101 which indicates she has a bachelor's degree.

Next let's consider the last four individuals in Table 1.1. They are all in household 68781 (a household we meet again in Chapter 5, Question 2). This is a household with two parents and two children. Both are 39 years of age and they have two children aged two and three, a boy and a girl. So far this looks like a prototypical American household, however, closer inspection reveals the parents are not citizens; we know this because the value of the variable CITIZEN equals 3, which indicates someone is not a citizen of the USA. The mom of this household does not work (we know this because there is zero value of UHRSWORK). The husband has a doctoral degree and his occupation is 055 (Electrical engineer). This person has the highest level of both education and income among the twelve individuals in Table 1.1.

Although not indicated in Table 1.1, the ACS also reports the place of birth for all respondents. It turns out the parents in household 68781 were born in Korea. The children were born here but are growing up with parents who cannot vote in U.S. elections. As they get older they will be strongly influenced by the community around them. Their interactions in schools, churches, and markets will strongly shape their national

self-identification. Thanks to the ACS, we know a lot about this family, but of course quantitative measures like the variables in the ACS can only tell us so much.

This book is not a memoir, but I have read some memoirs while writing it. A memoir can be thought of as a source of qualitative information, which contrasts with the quantitative nature of the ACS data. In his memoir, *Fresh Off the Boat*, which was later adapted into a popular television show of the same name, chef and restauranteur Eddie Huang describes growing up in a Taiwanese immigrant family in Orlando, Florida in the 1980s, developing tastes for both hip-hop and other elements of American culture, and the culture of his ancestors, especially food. Unlike Huang and his brothers, who were born in Taiwan, the kids in household 68781 are native-born Americans of Korean immigrant parents. There are other important differences between these two households. But it is possible some of the experiences Eddie Huang shares with his readers may be similar to experiences the kids in 68781 will have.

For example, one of Huang's stories relates to his embarrassment in bringing Taiwanese food to lunch at his largely white U.S. elementary school, and another to how unprepared he was for the texture when he finally tried typical American kids cuisine like macaroni and cheese and tuna fish sandwiches. Given they live in Silicon Valley, the kids in household 68781 will likely attend more diverse schools than did Huang, but these kids will be introduced to many types of American food outside the household, as were Huang and his brothers. The cultures from which Americans descend infuse with American culture as well, and today Eddie Huang owns a restaurant serving Taiwanese food in the trendy East Village neighborhood of New York City.

With a memoir like *Fresh off the Boat*, a reader learns intimate details about a family compared to the ACS where we have dozens of variables, but from a snapshot in time only. With a memoir, we know details about the extended family, where they have lived and what they have done throughout their lives. We cannot hope to gather this much detail from a 15-page survey that we expect millions of households to complete. Thus in this book I'll draw from diverse sources for the stories that will give life or at least help us imagine what the households in our data may really be like.

When interpreting raw data like those in Table 1.1, best practice calls for carefully consulting the codebook. This is a document that gives the meaning of every possible value of each variable. It's therefore worth repeating that a codebook for all variables discussed in this book appears in Table A.4

in the Appendix. Sometimes the value of a variable seems self-explanatory and you do not need to consult the codebook to determine its meaning; for example, when AGE equals 57 it seems obvious this means the individual is 57 years old. And in this case it is true. But what about when SEX equals 2? You might reason that because female comes before male in alphabetical order, 1 would represent a female and 2 a male, but if you consult the ACS codebook you'll find it is the opposite.

The RELATED variable can at times leave us guessing as to the real connections between people in a household. For example in household 67481, there is a father who lives with his seven-year old son, and a 27-year-old female. Her value of RELATED is 1115 and this indicates she is a "Housemate or roommate." You can see this is the eleventh option in the relationship question shown in Fig. 1.1. We can only imagine the real story about this 27-year-old lady who lives with these guys. Perhaps she rents a room from the family and interacts very little with them, or perhaps she is in a romantic relationship with the father and has a parental relationship with the boy. There is an option on the Census form to indicate someone is an unmarried partner, but this is not the information the father provided to the Census interviewer.

There are only 12 individuals and nine variables reported in Table 1.1, but the full ACS sample, starting from 2000 and, as of this writing, until 2018, includes data on almost 50 million individuals. From the 15-page survey form, the Census produces and distributes over 129 person variables. In addition, we know things about the households themselves, and the Census produces over 102 household variables.[1] Some examples of things we know about each household are illustrated in Table 1.2.

Table 1.2 tells us the total household income from all sources of all members of the household (HHINCOME) and the number of rooms in the home (ROOMS). For rented homes we know the monthly rent (RENT), and for owned homes the estimated current market value (VALUEH). Table 1.2 also indicates how many vehicles (VEHICLES) are available to members of the household. Take the first two households (67205 and

[1] IPUMS also constructs its own variables, such as those describing family interrelationship. I have written a blog post that contains a link to all the person and household variables the Census produced in 2015, for PUMA 068511. The variable names and coding values that appear in this book reflect IPUMS conventions, but it can be illustrative to see how the same data looks when it is distributed by Census.gov. http://mattholian.blogspot.com/2019/09/downloading-census-micro-data-ipums-or.html.

Table 1.2 ACS Raw Household Data, 2015, PUMA 068511

SERIAL	*HHINCOME*	*ROOMS*	*RENT*	*VALUEH*	*VEHICLES*
67205	93000	6	0	500000	3
67359	190000	7	0	800000	3
67383	19000	4	1200	9999999	1
67481	108000	5	2200	9999999	2
68781	115300	5	2400	9999999	2

Notes Data is 2015 ACS microdata for five households in PUMA 11 in Santa Clara County, California. The Data section of the file `script1.R` on the book's webpage produces a table with these data in it

67359). Both of these households own their homes. The ACS asks homeowners to estimate the value of their home if it were for sale today. San Jose is in one of the most expensive real estate markets in the country, and in 2015, these two households estimated their home's value as $500,000 and $800,000, respectively. Each of these households also happens to have three cars. The other three households are renters and reported rent ranges from $1200 to $2400. Note that these three households have a value of 9999999 for VALUEH; this means they are renters, not that their homes are worth just shy of $10 million. This example highlights another reason why reading the codebook is important. Finally, what about income? Table 1.1 indicated the wage income of each individual, but Table 1.2 presents the sum of all income earned by all people in the household. Thus Table 1.2 reports a household income of $93,000 for household 67205, which is just the sum of $72,000 and $21,000 reported in Table 1.1 for the two individuals who live in household 67205. Some of the other households earn income from other sources beyond wages (such as investment earnings).

Like the "person" variables in Table 1.1, the "household" variables shown in Table 1.2 are only a small subset of those available in the full sample. Clearly, the ACS data tells us a lot about the people and households it surveys. The information may not be complete, as in cases when investment income is mistakenly not reported, and it is likely that some people are hesitant to provide truthful information to the Census Bureau about somethings, as in the case of illicitly earned income. And perhaps a respondent is uncertain how to answer some questions (Is my lady friend who just moved in a partner or a roommate?) It may be hard to estimate what your home would sell for today if you haven't been following real estate market trends. Despite these limitations, the data provides a wealth of information about Americans.

The individuals and households in Tables 1.1 and 1.2 all live in the same geographic area. To protect the confidentiality of respondents, the Census Bureau does not share detailed information on the household's location with the public. Instead, they report the Public Use Microdata Area or PUMA in which the household lives. PUMAs are designed to contain about 100,000 people. So in a dense city like San Francisco PUMAs will be relatively small geographic areas, around the size of a zip code, but in rural areas a PUMA may be larger than a county.

Figure 1.2 shows a map of the USA that is divided into 2351 PUMAs.[2] Figure 1.3 shows detail of the San Francisco Bay Area. The PUMAs in the part of San Francisco where I live are somewhat smaller in land area than the PUMAs near where I work in downtown San Jose. This is a result of the fact that San Francisco has higher population density, and PUMAs are designed to hold about 100,000 people. In New York City there is a seven square mile area that contains more people than in the entire state of Wyoming.

Table 1.3 presents some statistics that I calculated with the microdata.[3] The file `script1.R` on the book's webpage carries out the calculations using ACS data from 2015. These statistics highlight some dramatic differences across seven illustrative PUMAs. In New York City, in one small section of Manhattan, the large majority (86.5%) of adults aged 25 and over hold bachelor's degrees. In the eastern portion of the city of Cleveland, Ohio (which is in central Cuyahoga county) this figure is only 15.4%.

The Manhattan neighborhood also stands out as having few children; only 6.3% of the people are under 18 years old, while this fraction is 20% or higher in some of the more suburban areas. It may be that living in a

[2] These maps were created using an open-source software program called QGIS. This book is in part a guide to R software, but R is not the only "open-source" software I used in writing this book. In Appendix A, I discuss a third software program called LaTeX which I used for word processing. Here, I wanted to provide some guidance on cartography. Making maps with computer software is part of a field called Geographic Information Systems (GIS). An excellent commercial version of GIS software is ArcGIS, but the free, open-source QGIS program is also easy to use. You can download the software and find training manuals at: www.qgis.org. More resources are at www.qgistutorials.com. Once you have GIS software, you need input files known as "shapefiles" or map "layers." Download these here: https://usa.ipums.org/usa/volii/tgeotools.shtml.

[3] It is possible to download some of these statistics from data.census.gov. While this is a good place to find commonly used statistics like average income, you won't find specialized statistics, like average lawyer earnings by college major. To find these a user has to calculate them themselves with the public microdata as I have done here.

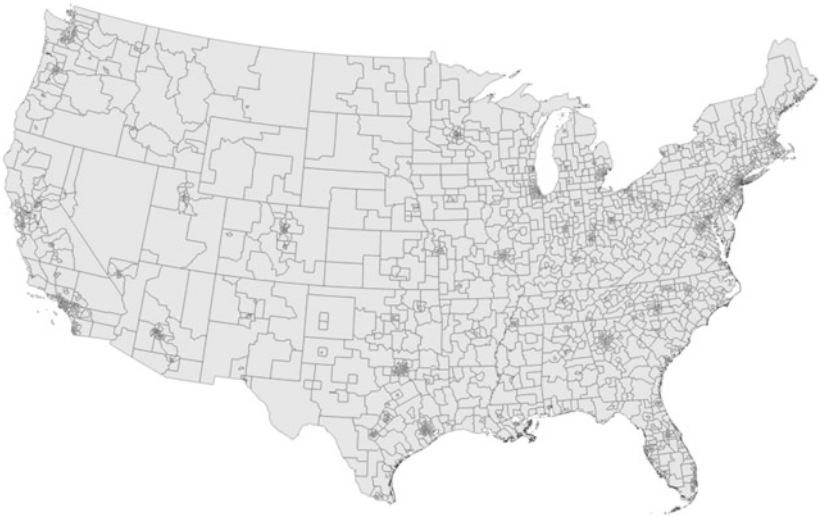


Fig. 1.2 Map of USA showing 2378 Public-Use Microdata Areas

high-density area like Manhattan causes families to have fewer children. It could also be that people with smaller families are more likely to live in these areas. This is a question of "selection versus treatment" that is a reoccurring theme in this book.

The final column in Table 1.3 does not actually contain a statistic (a number calculated with data) but rather a measure of the land area of each PUMA in square miles. In fact, this land area measure is not part of the ACS data. Of course, the ACS won't always contain all the measures we need, and sometimes we have to merge data from other sources on to the ACS. I discuss merged data in more detail in Appendix A. You can see in Table 1.3 that the PUMAs in New York City and, to a lesser extent the one in San Francisco, are much smaller than the other PUMAs, especially those in Ohio. Figure 1.4 shows New York City, highlighting two neighborhoods: PUMA 3603808 (Murray Hill) which includes the East Village neighborhood where Eddie Huang's restaurant is located, and Greenwich Village, the neighborhood where twentieth-century urbanist Jane Jacobs lived. She wrote passionately about the vibrant street life in her neighborhood in the 1950s when the so-called "urban renewal" policies were trying to remake cities better suited to automobile travel. Jacobs saw

Fig. 1.3 Map of San Francisco Bay Area, showing Public Use Microdata Areas, and indicating author's home and work locations

urban renewal as killing the urban vibrancy she loved, and indeed many U.S. cities went through a period of decline in the second half of the twentieth century. A section of Chapter 3 on migration describes research using the ACS that is related to the contemporary "back to the cities" movement.

This book is about how researchers use the ACS microdata, conveniently distributed by IPUMS, to calculate their own statistics. Many of the statistics we will see in the pages that follow will be averages or percentages like those in Table 1.3. It is also common to talk about differences between two

Table 1.3 Statistics for seven select PUMAs calculated using ACS microdata from 2015

Geography	*PUMA ID*	% *married*	%. *under 18*	% *college*	% *black*	% *white*	*Land area (sq. miles)*
NYC-Manhattan (Murray Hill)	3603808	40.4	6.3	86.5	1.7	77.9	1.6
San Francisco (North and West)	0607501	43.6	12.2	66.3	5.6	57.5	9.6
San Jose City (South Central)	0608511	68.0	24.9	47.4	4.0	58.2	12.8
Cuyahoga County, Ohio (West)	3900901	63.8	22.1	42.8	1.3	94.4	45.9
Cleveland, Ohio (East)	3900908	29.4	20.0	15.4	83.7	12.0	30.9
New Orleans City (Central)	2202401	30.8	16.7	38.0	56.8	39.4	16.6
District of Columbia (Central)	1100105	32.6	6.6	81.7	18.8	67.2	9.8

Notes All statistics were calculated with 2015 ACS microdata. The file `script1.R` on the book's webpage documents the analysis that produces these statistics. Persons in group quarters are excluded from the calculations. The statistic % married is the percent of adults age 25 and older that are married; % under 18 is the percent of persons under age 18; % college is the percent of adults age 25 and older with a bachelor's degree or higher; % black and % white are the percent of persons reporting race as black and white, respectively. Land area is the area in square miles of the PUMA, obtained from the Census Bureau's Gazetteer Files

averages. An average is a statistic, and the difference between two averages is also statistic. Most of the statistics reported in this book are nothing more than averages, and differences of averages of one sort or another.

Let's turn now to the first of many scholarly studies we will encounter in this book which uses the ACS data to calculate statistics. In an article titled, "Is economics a good major for future lawyers? Evidence from earnings data," the economist John Winters reported average lawyer earnings by college major, for the 25 most popular majors. This article was published in *The Journal of Economic Education*, a scholarly journal which is distinguished from popular press magazines, in addition to the nature of the material, by the fact that the articles go through a blind peer-review process. This study reveals that lawyers who were economics majors do quite well, with average annual earnings of $182,359. This is a key statistic from the Winters (2016) study that is replicated in the R and Stata files associated with this chapter. Among the 25 most popular majors, only electrical engineering majors do better, with average annual earnings of $219,383. The difference between these two averages (which are means not medians) is $37,024. (219,383–182,359=37,024). Therefore $37,024 is an example of a *difference in means*.

Fig. 1.4 Map of New York City Area, showing Public Use Microdata Areas, and indicating Greenwich Village and Murray Hill locations

How should we interpret these means? They are best thought of as descriptive statistics, measures of what the world looks like, not explanations for why or how it is. But do these figures suggest an economics student could really make $37,024 more if they switched to electrical engineering? In other words, does studying engineering cause you to earn more, all else equal? This figure reflects what empirical economists call both *selection bias* and a *treatment effect*. Selection bias results when students who major in electrical engineering in college are likely to earn more later in life for reasons apart from the curriculum they studied. Perhaps they're smarter, or more willing to sacrifice leisure for higher income. The effect of the curriculum itself on earnings, which is what a student considering switching

majors wants to know, is the treatment effect.[4] At this point, most students can't change who they are, but they do have to decide what to major in.

The use of the terms selection and treatment is based on an analogy to *randomized experiments*. In these types of studies, treatment is administered by the experimenter randomly, for example, by the flip of a coin. In this lawyer earnings example, there is no experimenter randomly assigning students into economics or electrical engineering majors. We could imagine some process whereby universities do have the ability to assign students to majors, but, most of the time this sort of experiment would be prevented by the ethnical consideration that, in a free society individuals should be able to select their college major themselves. It is often still helpful to consider what a hypothetical *ideal experiment* would look like, as this forces us to define the causal effect we are estimating.

A good way to understand why a difference in means estimated with the ACS microdata is often not a good estimate of a treatment effect is to recognize that data from the ACS are an example of *observational data*, not *experimental data*. Experimental data is produced through conducting a formal experiment with random assignment of subjects into control and treatment groups. This contrasts with survey data, which is produced through questionnaires by randomly sampling households in a population. Observational data is data produced in non-experimental settings, for example, by looking around (observing) and recording what we see. There is no experimental manipulation, only recording facts as they are seen or reported. The Census Bureau asks respondents to answer questions, but in principle some of the questions could be answered by the interviewer just by observing, such as how many average size rooms are in a house, how many cars are visible, and so on.

With observational data, a difference in means like our figure of $37,024 will usually be a mix of both selection and treatment effects. At this point we can't say much about the future earnings a student can expect by switching to electrical engineering from an economics major. One thing we can say

[4] Angrist and Pischke (2014, p. 10) discuss selection bias in the context of an equation, which I adopt as a definition of the term: Selection Bias = (Difference in Means) − (Average Causal Effect). Here, selection bias is the entire gap between what we observe ($37,024, the difference in means) and the true impact of the treatment, which generally is unknown, but could be measured in an ideal randomized experiment. There's a lot of jargon in econometrics, some of it unfortunate, and some of it necessary to discuss nuanced concepts. Take, for example, the treatment effect. The effect of the economics curriculum likely varies across people. The average of the individual effects is known as the *average causal effect*.

is that this figure of $37,024 is a description of the population of working lawyers. It is best thought of as a *descriptive statistic*, rather than a causal effect (i.e., the treatment effect in an ideal experiment that randomly assigns college students to an electrical engineering versus economics major). In the ACS data, people are not assigned to majors and instead select what they major in, based on their own innate preferences and abilities. If, hypothetically, people were randomly assigned to majors, then differences in average earnings could be interpreted as causal effects, but here they cannot.

A theme of this book is understanding when observational data can and cannot be used to estimate causal effects. The example above illustrates when statistics cannot be interpreted as a causal effect. Below, I introduce two techniques that do allow us to estimate causal effects with observational data. But before continuing I want to clarify an important point: The average earning estimates of $182,359 and $219,383 are valuable measures of reality. Just because they are descriptive statistics and not treatment effects or causal effects does not mean they are not valuable contributions to the body of knowledge. There seems to be some prejudice in economics and other social sciences against descriptive research. Political scientist Justin Grimmer (2015) has argued that, "Political scientists prioritize causal inference...often pejoratively dismissing measurement...as 'mere description' ...." He points to research by Gerring (2012) that showed 80% of articles published in *American Political Science Review* focus on causal inference. In light of this, the important point I want to make is that even though much of this book focuses on econometric techniques for causal inference, it certainly does not aim to dismiss the importance of descriptive research. This lawyer earnings case study, the PUMA-level statistics presented in Table 1.3, and many other descriptive statistics to be discussed in chapters that follow highlight the valuable role the ACS plays in measuring social quantities.

Often, as in the case of policy evaluation and CBA, we do need precise estimates of causal effects as well as accurate descriptive measures of social quantities. Luckily we have techniques to estimate causal effects with observational data. One is called *regression control*, and we could reanalyze the lawyer earnings example if we had data on factors that jointly determine a person's choice of major and their earnings. Intelligence is one such factor, and test scores are a possible but controversial measure of intelligence. The second technique encompasses several types of research designs, which share a common search for *natural experiments*. Natural experiments are settings where social or political processes end up assigning treatment in

a way that is as if it were randomly assigned by an experimenter in a true experiment. When we find a natural experiment, we don't need data on control variables, and sometimes a difference in means that we estimate with observational data can be interpreted as a causal effect.

As a hypothetical example of a natural experiment, imagine a university where the two most popular majors are electrical engineering and economics, and the number of students who wish to major in them was greater than the number of spaces available. So the university creates a wait list and assigns students on the wait list into majors based on a lottery. In this case, the treatment—whether a student is an economics or electrical engineering major, is as good as randomly assigned, and a comparison of earnings by these majors five years post-graduation can be interpreted as an average causal effect of the major itself on earnings. This would be an example of a natural experiment if university administrators at the time didn't intend to randomly assign major to study the causal effect of curriculum on earnings.[5]

Like a lot of econometrics jargon, the term natural experiment doesn't always clearly convey its meaning. After all, there's not much natural about the idea of college administrators assigning students to majors through a lottery. What makes the example a natural experiment is that the treatment was assigned randomly by some process other than a true experiment.

Sometimes with natural experiments, the process *is* natural, as with weather shocks. As another example of a natural experiment where the term natural makes more sense, the gender of a child can be said to be determined by nature. If child gender is as good as randomly assigned, even when it's not actually randomly assigned by an experimenter, we can use basic statistics such as a difference in means to provide a compelling estimate of a causal effect. We'll see an example of this at the end of this chapter.

[5] If their intention behind randomly assigning students to major was to study the effect of major on income, we would call it an actual experiment (or a randomized experiment, or maybe a field experiment) but not a natural experiment. See Dunning (2012) for further discussion of natural experiments. Bleemer and Mehta (2021) use a grade point average policy at UC Santa Cruz, in a technique called *regression discontinuity*, to study the causal effect of the economics curriculum on earnings.

Another unfortunate econometric term is *regression*.[6] Despite what the name suggests, it actually refers to a widely used statistical technique to estimate empirical relationships. To make the idea of regression precise, fitting a trend line to a scatterplot is one application of regression (Fig. 1.5 in this chapter's Review Question 7 illustrates this). Regression can do much more than fit trend lines in two-dimensional figures, however, and is widely used in all of the studies I discuss in this book; therefore, I now turn to introduce regression techniques.

We begin with the simplest form of regression, known as *bivariate regression*, because it involves two variables. Staying with the lawyer earnings example, consider the following equation:

$$INCEARN_i = \alpha^s + \beta^s ECON_i + e_i.$$

On the left-hand side, $INCEARN_i$ is a variable equal to reported annual earnings for individual i, and we refer to this variable as either the left-hand side variable, or the *dependent variable*. This model has one *independent variable* on the right-hand side, $ECON_i$ which is a *binary variable*, meaning it takes on values of zero or one; it is equal to one if individual i was either an economics or business economics major, and zero if they were any other major. Independent variables can be *continuous* in regression models, though here it is binary, and likewise, there is nothing preventing the dependent variable from being binary. The individuals in the sample are what I refer to as the *estimation subsample*. To estimate the model shown above, I use data on: persons in occupation code 178 (lawyers), with a professional or doctoral degree, who are between the ages of 30 and 61, and were sampled in the ACS between 2009 and 2013.[7]

[6] In 1886, Francis Galton found that children of very tall parents tend to be shorter than their parents, and he described this as, "regression to the mean." The statistical technique he developed to study this phenomenon used an equation that has since become known as a "regression equation." See also Bailey (2017, Chapter 3, footnote 2) and Angrist and Pischke (2014, pp. 79–81) on the history of the term regression.

[7] Sometimes the description of the estimation subsample is referred to as a model's "data rules." Note here the estimation subsample includes persons with all undergraduate majors, not just economics and electrical engineering. Determining the estimation subsample a researcher used is often a major challenge in replicating a study, but it is the critical first step. In the file `script2.R` on this book's webpage, one line of code defines the estimation subsample for the Winters (2016) replication. This line creates a data set (or "data frame" in R language) that I named "subset2w": `subset2w = subset(ACSmaster, OCC1990==178 & EDUCD>114 &`

In addition to having two variables, there are also two *regression coefficients* in this model, α^s which is the *constant* (also referred to as the intercept) and the coefficient on $ECON_i$, β^s (also referred to as the slope). In this book, I always use Greek letters to refer to coefficients. The point of regression is to find estimated values of these coefficients. The final term in the equation above is called the *error term*. Intuitively, the error term explains everything about an individual's choice that is not explained by the independent variable. Mathematically, the error term is something to be minimized. Regression estimates values for the α^s and β^s coefficients, in a technique called *ordinary least squares*. While we never observe the error term, the estimated value of it is called a *residual*. The residual is the difference between the actual value of earnings for lawyer i and the predicted or *fitted value*.

After collecting data on the dependent and independent variables, we are ready to "run the regression," which in this example means, asking our statistical software package to estimate values of the α^s and β^s that best predict someone's earnings, based on whether or not they were an economics major. Regression coefficients are statistics because they are calculated with data. It may be intuitive that, if all you know about a lawyer is whether or not they were an economics major, the best you can do in predicting their earnings is use the average earnings of all lawyers who were economics majors. Likewise, if all you know about someone is that they were not an economics major, the best prediction, by the ordinary least squares criteria, is the average earnings of all lawyers that were not economics majors. I estimate the equation above in the file `script2.R` that replicates the Winters (2016) study. The results are reported below:

$$INCEARN_i = 149{,}709 + 32{,}650 \times ECON_i$$

`AGE>29 & AGE<62 & YEAR>2008 & YEAR<2014)`. Here, `subset2w` is the name I gave the data frame which is the estimation subsample, and `subset()` is an R function that creates a smaller data frame from a larger data frame. The larger data frame, `ACSmaster`, is the IPUMS extract with 61 variables from 14 survey years that I discuss in Appendix A, and `subset2w` is a much smaller data frame that only contains lawyers surveyed in certain years with certain other characteristics. The data frame `ACSmaster` is a large file. It is nowhere near as large as the file would be if the extract included all variables and all samples available from IPUMS, but it is large enough to enable me to estimate every statistic I present in this book. Every statistic presented in this book is estimated on data that is a subset of the `ACSmaster` data frame described in Appendix A.

where the estimate of α^s is 149,709 and the estimate of β^S is 32,650. (We don't get an estimate of the error term. Statistics based on the residuals do contain useful information for making inferences, but they are not a focus here.)

We can use this model to predict by calculating *fitted values*. Fitted values are predictions from the model. I was an economics major in college and I became a college professor. If I instead would have become a lawyer, we could use this model to predict my income; to find my predicted earnings, we substitute a value of one into the equation for $ECON_i$, and solve: 149,709 + 32,650(1) = 182,359. This is the average earnings of lawyers who were economics majors, and is the exact same statistic we saw before and which was reported in the Winters (2016) study.

My friend Dave, who I grew up with in PUMA 3900901, did become a lawyer but he was a marketing major not an economics major. To find his predicted earnings using this equation, we substitute a value of zero into the equation for $ECON_i$, and solve: 149,709 + 32,650(0) = 149,709. This is another fitted value, and it happens to be the average earnings of lawyers who were not economics majors. This model thus predicts Dave earns $149,709. Would he be earning $32,650 more if he would have just chosen an economics major? As discussed above, $32,650 is best thought of as a descriptive statistic rather than a treatment or causal effect. It seems far fetched to think that if he would have taken just a few different classes to be a business economics major, his income would be 20% higher today.

Thus, $32,650 is the difference in means between economics and all other majors. It is a biased estimate of the causal effect of the economics college curriculum on later-life earnings, because it was calculated with observational data—students select their major themselves and are not randomly assigned into majors by an experimenter. If we can't actually carry out the ideal experiment (which is the normal state of affairs) we may at least be able to estimate a less-biased version of the causal effect, using the regression control technique I introduced above.

Regression control is a statistical technique to solve or at least reduce selection bias. It requires adding *control variables* to the right-hand side of the regression equation. A regression equation with more than one independent variable is referred to as *multivariate regression* in contrast with the bivariate regression equation we saw above. As it turns out, men are more likely to major in economics than women. Among lawyers with an economics major in our sample, 2093 were male and only 664 were female. Male lawyers also earn more than female lawyers, on average. Mean

female lawyer earnings are $117,330 while mean male lawyer earnings are $172,438. Thus, part of the reason we find economics majors earn more than other majors is because there are more males in this major, and males earn more money for reasons apart from their training (possible reasons for this include discrimination, and the fact historically women have taken on a larger share of household responsibilities).

A multivariate regression equation designed to "control" for the gender selection effect is:

$$INCEARN_i = \alpha^l + \beta^l ECON_i + \gamma FEMALE_i + e_i^l.$$

This model has two independent variables on the right-hand side, $ECON_i$ as before, and one new one, $FEMALE_i$, which is a binary variable equal to one if the respondent is female and zero if male. I use superscripts on the coefficients α^l and β^l to indicate that these are different from those in the bivariate regression equation above. Estimating this equation we find:

$$INCEARN_i = 170{,}457 + 25{,}206 \times ECON_i - 54{,}207 \times FEMALE_i.$$

The key thing to note here is the coefficient on $ECON_i$ is now 25,206, which is less than the estimated coefficient in the bivariate regression above where it was 32,650. Thus, controlling for gender, we find a smaller economics major effect.

It is tempting at this point to think we can just keep adding variables to the right-hand side that explain the dependent variable. Building elaborate multivariate regression models is often a part of the regression control strategy. But what is really important is to add independent variables that meet the following two conditions: (1) explain the dependent variable, and (2) are correlated with the main independent *variable of interest*. These are the two OVB conditions, and failing to include variables that meet both conditions leads to *omitted variable bias* (OVB).

One of the limitations of regression control is that we usually don't have data on all the factors that cause bias. For example, we can't measure ambition or how much someone values the intrinsic versus financial rewards in any particular area of law. But, we can at least argue $25,206 not $32,650 is a more realistic upper bound estimate of the treatment effect of majoring in economics for aspiring lawyers, and this example illustrates the promise

of the regression control technique in reducing if not solving completely the problem of selection bias.[8]

Before concluding this chapter, let's consider one more regression example. This time, we can interpret the coefficients as causal effects, even though they were estimated using observational data. In the example, treatment is not randomly assigned by an experimenter, but it is randomly assigned by nature. The question we aim to answer is, do families with same gender children live in homes with fewer bedrooms? We would expect they would, if it is easier for children of the same gender to share a bedroom. Consider the regression equation shown below, which I estimate on an estimation subsample of 458,837 married-couple households with precisely two children:

$$BEDROOMS_i = \alpha_0 + \alpha_1 SAMESEX_i + e_i$$

where the dependent variable $BEDROOMS_i$ is equal to the number of bedrooms in household i's home, and the independent variable $SAMESEX_i$ is a binary variable equal to one in boy–boy and girl–girl households, and zero in boy–girl households. In the notation here, α_0 is the constant (or intercept), and α_1 is the slope (or the coefficient on the variable $SAMESEX_i$).

The file `script3.R` estimates this model. I find the estimate of α_0 is 4.4. This is because the average number of bedrooms is 4.4 in households where the children have different genders (i.e., one boy and one girl). The estimate of α_1 is -0.03; the average number of bedrooms is slightly less in households with same gender children. Adding the coefficients together, 4.4 – 0.03 = 4.37, and this is the average number of bedrooms in households with same gender children.

[8] Selection bias was defined in footnote 4 as the entire gap between the difference in means and the true average causal effect. In the context of regression control, we often discuss the different concept OVB, which is the gap between estimated values of the short and long coefficients. Specifically here, OVB = \$32,650 − \$25,206. Unless our long regression includes all possible factors that meet the two OVB conditions (they influence the dependent variable and are correlated with the independent variable of interest), OVB will be less in magnitude than selection bias. Review Question 8 presents an interesting relationship between coefficients from various regressions and the two OVB conditions.

In terms of magnitude this is a small effect,[9] but because of the large sample size, it is highly statistically significant—we would be very likely to again find a small, negative effect if we estimated this same model on a different sample drawn from the same population (say, if we used a sample of ACS data from a different survey year). Statistical significance is a question of *inferential statistics*, which is important but is not a major focus in this book.

Is this estimate of α_1 equal to –0.03 compelling evidence that a child's gender affects their parents' housing decisions? This is a question of causal inference. My answer is yes. The gender of children is, almost always, as good as randomly assigned. This means, for example, we would expect to find (and in the analysis, we do find) that boy–boy and girl–girl households have the same income, on average, as different child gender households (in other words, the difference in mean income is statistically insignificant across these two household types). These two types of households are the same in almost all other ways too, at least on average, except for the fact that half of them were assigned by nature to have children of the same gender, while the other half were not. Thus I find this to be compelling evidence that child gender is a causal factor in family housing decisions.

This discussion of child gender and bedrooms may not seem like one of the critical social issues of our time. However, in our study of transportation in Chapter 6, we'll see the interplay between child gender and housing outcomes can actually help us find solutions to the pressing issue of climate change (Holian 2020).

We have covered a lot of ground in this chapter. Navigating econometrics jargon and concepts takes getting used to. I've introduced dozens of terms because econometrics has its own language; I wrote the glossary to this book to help beginners here. The main takeaway from the chapter is that regression is a tool for calculating means, differences in means and coefficients in multivariate models. Sometimes those statistics have causal interpretations, and sometimes they are "merely" descriptive statistics. This chapter also introduced the technique of regression control, and the next chapter on homes presents a detailed case study of this technique.

[9] The gender of a child does seem to be a more important factor for households on the margin between three- and four-bedroom homes. Analysis carried out in the file `script3.R` finds that 54% of households in three-bedroom homes are same-gender child households, while only 47% of households in four-bedroom homes have children of the same gender.

Key Terms

Variable	Codebook	Difference in means
Selection bias	Treatment effect	Average causal effect
Randomized experiment	Ideal experiment	Observational data
Experimental data	Descriptive statistics	Regression control
Natural experiment	Regression	Bivariate regression
Dependent variable	Independent variable	Binary variable
Control variable	Estimation subsample	Regression coefficients
Constant	Error term	Residual
Fitted value	Ordinary Least Squares (OLS)	Multivariate regression
Variable of interest	Inferential statistics	Omitted Variable Bias (OVB)
Random sampling	Sample weights	Linear Probability Model (LPM)
Regression discontinuity		

Questions for Review

1. Compare and contrast the numbers in Tables 1.1 and 1.3 and identify them as data and statistics. Which organization collects the ACS data, which organization specializes in distributing the ACS data, and what is the role of researchers?
2. Evaluate whether the statistics related to bedrooms and sibling gender better reflects causal inference or descriptive statistics.
3. Describe in words the estimation subsamples for the lawyer earnings model presented in the chapter, by translating the R code from footnote 8.
4. How would someone conducting a randomized experiment which assigned subjects into treatment and control groups use a bivariate regression model with a binary independent variable? Is it correct to say someone running a regression with observational data, where the data was collected through a survey that used simple *random sampling*, is carrying out an experiment? Is observational or experimental data used in the analysis of a natural experiment?

Questions 5–8 are Empirical Exercises and use `script4.R`

5. Table 1.4 contains data from six individuals that are lawyers who studied either business economics or marketing in college, and who were surveyed in the ACS between 2009 and 2012. Calculate the mean of INCEARN for economics majors and marketing majors separately, using a paper and pencil. After trying it by hand, run `script4.R`.

Section 3a of this script calculates these two group means (and illustrates commonly used statistical procedures not directly related to this question, including: summary statistics, frequency tables, scatterplots, difference in means hypothesis tests, and regression). Report your findings. Do you get the same results whether using a pencil or statistical software? Finally, discuss whether the difference in mean earnings you find here between economics and marketing majors represents a causal effect or if it suffers from selection bias, and why.

6. This question illustrates two best practices in analysis of ACS data concerning inflation adjustments and sample weighting.

   (a) Read the codebook description for the CPI99 variable in Appendix A. Then add a new variable to Table 1.4, named INCEARNadj, which expresses earnings in constant 1999 dollars. Recalculate mean earnings for economics and marketing majors, using a paper and pencil and the inflation-adjusted earnings variable. Next, express earnings in constant 2015 dollars and recalculate the means (*Hint:* the price level rose 43% between 1999 and 2015, so multiply the 1999 figures by 1.43).
   (b) Simple random sampling is often used to collect data that is representative of a population, but even in large samples like the ACS, sometimes *sample weights* can be used to make the sample more representative of the population from which it was drawn. The PERWT variable indicates how many persons in the U.S. population are represented by a given person. Recalculate mean earnings for economics and marketing majors (you may use INCEARN in current dollars, or in constant dollars using a base year of your choice). Hint: imagine there are 86 people like the first observation in your sample, 63 like the second, and so on.
   (c) Section 3b of `script4.R` illustrates use of the CPI99 and PERWT variables. Does the software produce the same results you found in parts a and b?
   (d) How does a failure to adjust for inflation affect the results? How does a failure to apply sampling weights affect the results?

7. Section 3 of the R script analyzes a full subset of data from the 2009–2013 ACS, of lawyers between the ages of 30 and 61, from which the six observations in Table 1.4 were taken. I find there

Table 1.4 Select observations from 2009 to 2012 ACS, Lawyers with two majors

	SERIAL	*YEAR*	*INCEARN*	*ECON*	*CPI99*	*PERWT*
1	316038	2009	60000	0	0.777	86
2	838883	2009	80000	0	0.777	63
3	391901	2009	100000	0	0.777	79
4	1170872	2009	80000	1	0.777	62
5	177961	2012	90000	1	0.726	187
6	738723	2009	100000	1	0.777	143

Notes The last line of the Data section of the file `script4.R` on the book's webpage produces a table with these data in it

371 Business Marketing majors and 113 Business Economics majors were surveyed. Average inflation-adjusted earnings were $140,990 among Business Marketing majors and $179,582 among Business Economics majors. The difference in means is statistically significant ($p = 0.005$).

(a) Should these statistics be interpreted as causal effects, descriptive statistics, or something else? Why?
(b) Discuss an ideal randomized experiment that would measure the impact of studying business economics versus a marketing curriculum on earnings.
(c) Describe a hypothetical policy that assigns students to one of these two majors that is not a randomized experiment, but that approximates one.
(d) Examine the regression line shown in the scatterplot in Fig. 1.5, and find the equation of the line, using the statistics reported in this question (you can also find the estimated constant and slope coefficients in Column 1 of Table 1.5).

8. Table 1.5 contains estimates of the models discussed in the section of the chapter on lawyer earnings, but using an estimation subsample of lawyers that were either business economics or marketing majors. Tables like 1.5 are the conventional way regression results are reported, so reading published research requires understanding, at a minimum, how to identify estimated regression coefficients within tables like these. For now, focus on coefficient estimates and ignore all other numbers (you'll learn about the other numbers if you take

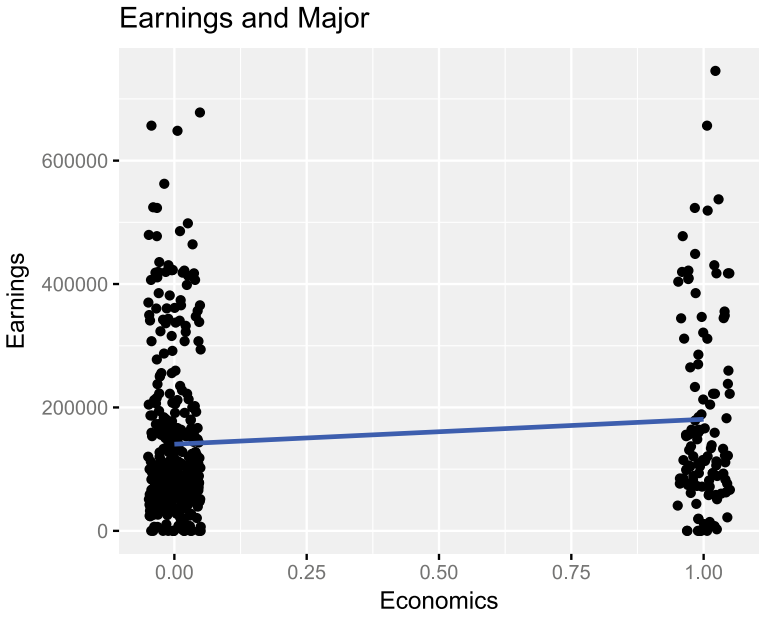


Fig. 1.5 Scatterplot showing annual earnings on y-axis among lawyers that majored in marketing (for whom $X = 0$) and economics (for whom $X = 1$)

an introductory econometrics class). Focus also on the results in columns 1–3 only (we will discuss column 4 in a review question at the end of Chapter 3).

(a) Using the "short" regression estimates reported in Table 1.5 column 1, identify the estimated values of α and β^s.
(b) Using the "long" regression estimates reported in Table 1.5 column 2, identify the estimated values of α^l, β^l and γ.
(c) Identify the estimated values of π_0 and π_1 in the auxiliary regression, reported in column 3.

Table 1.5 Regression results: lawyer earnings and the economics major

	Dependent variable:			
	INCEARNadj		*FEMALE*	*INCEARNadj*
	(1)	*(2)*	*(3)*	*(4)*
ECON	38,592**	28,866*	−0.1258**	27,417
	(16,158)	(15,705)	(0.0585)	(19,218)
FEMALE		−77,320***		−78,223***
		(11,937)		(12,804)
(ECON ∗FEMALE)				4550
				(33,446)
Constant	140,990***	173,416***	0.4194***	173,794***
	(7324)	(10,191)	(0.0297)	(10,751)
Observations	484	484	484	484
R^2	0.015	0.098	0.012	0.098
Residual Std. Error	1,262,321	1,208,934	4.7	1,210,162

Notes This table contains regression results for the sample of lawyers age 30–61, surveyed in ACS years 2009–2013, with professional or doctoral degrees, whose undergraduate major was either business economics or marketing. Earnings are measured in constant 2015 dollars, and weighted to reflect sampling probability. Heteroskedastic robust standard errors in parentheses, $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$
Regression models in columns:
(1) $\text{INCEARNadj}_i = \alpha + \beta^s ECON_i + e_i$
(2) $\text{INCEARNadj}_i = \alpha^l + \beta^l ECON_i + \gamma FEMALE_i + e_i^l$
(3) $\text{FEMALE}_i = \pi_0 + \pi_1 ECON_i + e_i$
(4) $\text{INCEARNadj}_i = \beta_0 + \beta_1 ECON_i + \beta_2 FEMALE_i + \beta_3 (ECON_i \times FEMALE_i) + e_i$
Variable descriptions: $INCEARNadj_i$ is earned income in 2015 dollars, $ECON_i$ is a binary variable equal to 1 for business economics and zero for marketing majors, and $FEMALE_i$ is a binary variable equal to one for females and zero for males

(d) Footnote 8 defined OVB as the difference between the short and long coefficients. Use the results in columns 1–3 of Table 1.5 to verify the omitted variable bias (OVB) equation: $\beta^s - \beta^l = \gamma \times \pi_1$.[10] Discuss how this equation helps us remember the two OVB conditions.

[10] The OVB equation requires estimating a so-called auxiliary regression, the equation for which is: $FEMALE_i = \pi_0 + \pi_1 ECON_i + u_i$. We call OLS models with binary dependent variables like this one *linear probability models*, and we interpret fitted values from them as predicted probabilities. The right-hand side of the OVB equation highlights that, if we omit a variable in a regression equation that is (1) highly correlated with the main variable of interest (we see this in the auxiliary regression on π_1) and (2) an important determinant of the dependent variable when it is included (we see this in the long regression on γ), the estimated coefficient in the bivariate regression will suffer from OVB. See also Angrist and Pischke (2014, p. 71) and Bailey (2017, Section 5.2).

References

Angrist, Joshua D., and Jörn-Steffen Pischke. Mastering 'metrics: The path from cause to effect. Princeton University Press, 2014.

Bailey, Michael A. Real econometrics: The right tools to answer important questions. Oxford University Press, 2017.

Bleemer, Zachary, and Aashish Mehta. "Will studying economics make you rich? A regression discontinuity analysis of the returns to college major." American Economic Journal: Applied Economics. Forthcoming, 2021.

Dunning, Thad. Natural experiments in the social sciences: A design-based approach. Cambridge University Press, 2012.

Gerring, John. "Mere description." British Journal of Political Science (2012): 721–746.

Grimmer, Justin. "We are all social scientists now: How big data, machine learning, and causal inference work together." PS, Political Science & Politics 48, no. 1 (2015): 80.

Holian, Matthew J. "The impact of urban form on vehicle ownership." Economics Letters 186 (2020): 108763.

Huang, Eddie. Fresh off the boat: A memoir. Spiegel & Grau, 2013.

Winters, J. V. "Is economics a good major for future lawyers? Evidence from earnings data." The Journal of Economic Education 47, no. 2 (2016): 187–191.

PART II

Regression Control

Learning Goals for Part II

1. Explain how comparing average electricity expenditures for homes built in different periods could provide a misleading picture of the relative energy efficiency of homes of different vintages.
2. Explain why a regression control approach that estimates period-of-construction effects along with controls variables like number of rooms in the home provides a more compelling estimate of the home's level of energy efficiency.
3. Define the following key terms: categorical variables, ordinal variables, logged variables, and fixed effects.
4. Critique the regression control model presented in this chapter by listing a relevant control variable that was not included in the model.
5. Describe the path a researcher can take from replicating a study, to extending it, to doing original research.

CHAPTER 2

At Home: Housing and Energy Use

For many Americans, owning a home is the most important part of the American Dream. The American Community Survey (ACS) asks all respondents whether they own or rent their home. Housing costs are the biggest item in the average U.S. household's budget (at 26%, according to data collected by the U.S. Bureau of Labor Statistics in their Consumer Expenditure Survey).[1] The ACS asks renters how much they pay in rent, and it asks owners to estimate the market value of the home if it were sold today.

The ACS also asks about household energy use, including the source (electricity, natural gas, etc.) of home heating fuel, and expenditures on each source. A household's expenditure on home energy use is not a huge fraction of their overall budget, but it is responsible for a large fraction of the carbon emissions they generate. Activities associated with home energy use (space heating, air conditioning, water heating, lighting, refrigeration) together are the source of approximately 14.9% of the average US household's carbon emissions (Nordhaus 2013, p. 161).

There are many important housing questions that can be studied with the ACS. The studies I discuss in this chapter deal with policies for reducing household energy use. We can reduce our carbon footprints by buying

[1] Unlike the Consumer Expenditure Survey, the ACS does not focus on all consumer expenditures, but the ACS does contain detailed questions about housing and household energy expenditures. https://www.bls.gov/news.release/cesan.nr0.htm.

M. J. Holian, *Data and the American Dream*,
https://doi.org/10.1007/978-3-030-64262-4_2

energy efficient refrigerators and other appliances, and also by designing our homes with more attention to details like insulation in the first place. Some U.S. cities, like Berkeley, have taken what may seem to some as drastic steps, including banning the use of natural gas fuel in new homes altogether. Across the state in California, new single-family homes are now required to install solar panels.[2]

The main case study reviewed in this chapter is by Dora Costa and Matt Kahn. It provides evidence on the possible effect California's building codes have played in the "greening" of the housing stock, by examining the relative energy efficiency of U.S. housing of different vintages. I discuss this study in detail because it is an excellent example of the use of *regression control* to estimate causal effects. I also discuss a study I carried out, which replicates the Costa and Kahn model, and also creates new knowledge on this topic. This chapter also clarifies what is meant by a replication, and illuminates the path a researcher can take, from replicating a study to carrying out original research.

Building Codes and Home Energy Use

In 1978, California adopted the nation's first building energy codes, which set minimum standards for various aspects of a building's design as it pertains to its energy use. Typically, building energy codes regulate specific design features of homes, such as the amount of insulation required behind walls and above ceilings, the thickness of window glazing, and so on. Today, most but not all states have followed California's lead and have adopted at least some form of building energy codes.

Making a home more energy efficient will not necessarily reduce energy use as much as an engineer might predict. One reason is an effect economists call "the rebound effect." As an example, the occupants of an energy efficient home may use the air conditioner (AC) more than they otherwise would. They may decide to bake a cake on a hot day, rather than postpone baking until the sun goes down. It can be time consuming to open and close all the windows in large homes; if the home is energy efficient, a household may decide to just keep the windows closed and use the AC all

[2] Berkeley becomes first U.S. city to ban natural gas in new homes. Sarah Ravani. July 21, 2019. San Francisco Chronicle. Fact Sheet. 2019 Building Energy Efficiency Standards. https://ww2.energy.ca.gov/title24/2019standards/documents/2018_Title_24_2019_Building_Standards_FAQ.pdf.

the time, whereas before they would have opened and closed them based on outside conditions.

The rebound effect accounts for human behavior and may mean energy consumption estimates we see on the technical specification sheets for building energy codes overstate the actual energy savings that will result from them. So when the California Energy Commission claims that, "Single-family homes built with the 2019 standards will use about 7 percent less energy due to energy efficiency measures versus those built under the 2016 standards," there are reasons to be skeptical. This is not to say the more stringent energy codes won't help, but the 7% figure (which was calculated by engineers based on a model home's technical specifications) may prove to be an over estimate.

With this background in mind, I now turn to the details of a 2012 electricity expenditure study by Dora Costa and Matt Kahn titled, "Electricity consumption and durable housing: understanding cohort effects." This study used data from the 2000 decennial Census long form. Although this is a book about studies that use the ACS data, I discuss this study in detail here, for a few reasons. First, the ACS was modeled on the decennial Census long-form survey, as discussed in Chapter 1, so the questions on the ACS are essentially identical to those on the 2000 long form. Second, I have carried out a replication and extension of the Costa and Kahn study that uses recent ACS data, which we'll review at the end of this chapter.[3]

The most basic way one could use these data to study the effect of energy codes is to examine energy expenditures across homes that were built in different decades or periods. The two key IPUMS variables required to do this are COSTELEC and BUILTYR. The first of these is the annual electricity cost for each household. This is a fairly straightforward variable.[4] Meanwhile, BUILTYR is an *ordinal variable* that ranges from 1 to 6 for the estimation subsample (which in Costa and Kahn was single-family homes

[3] There's a few other reasons the Costa and Kahn study is a good fit for this book. The data and code are all available on OpenICPSR. Second, it was published in the Papers and Proceedings edition of the *American Economic Review* (AER). This edition of the AER, now a standalone journal, contains only short articles, usually around five pages, which are the perfect length for beginning students (Elliott 2004).

[4] Like nearly all IPUMS variables, this one does need to be recoded before using it in analysis. A value of 0 means data on COSTELEC is not available, not that the household spent nothing on electricity. Also, 9993 means "No charge or no electricity used," and 9997 indicates, "Electricity included in rent." If these values are not recoded, these observations will be counted as having extremely high electricity bills when in fact they do not.

built between 1960 and 2000). Its value indicates the period in which the home was built. For example, a value of BUILTYR equal to 4 indicates the home was built in the 1980s (1980–1989), a 5 means it was built in the 1970s, and a 6 in the 1960s. For homes built in the 1990s, there are three categories: a 1 means the home was built in 1999, a 2 means it was built between 1995 and 1998, and finally homes built between 1990 and 1994 have a value of BUILTYR equal to 3.

Because BUILTYR is an ordinal variable (and similar to categorical variables, which have no intrinsic ordering) it would be inappropriate to regress COSTELEC on BUILTYR directly. Even though higher values of BUILTYR indicate the home is older, the periods of construction are unequal; some periods are decades while one period is just one year. Thus, rather than include the BUILTYR variable directly in a regression model, we create one or more binary variables based on it and use the binary variables in the regression model. This is illustrated in the equation below:

$$COSTELEC_i = \alpha_0 + \alpha_1 YB1980s_i + u_i,$$

where the dependent variable $COSTELEC_i$ is household i's annual electricity expenditures, and $YB1980s_i$ is a binary variable, equal to one if the home was built between 1980 and 1989, and zero if not. Recall, in the last chapter we saw how a bivariate regression model with a binary independent variable, like the one here, can be used to calculate the difference between two group averages (the difference in means). Estimating this model with the data shared by the authors on the *American Economic Review* webpage, I find:

$$COSTELEC_i = 1{,}109 + 20 \times YB1980s_i.$$

Like all models of this form, the estimate of the constant term, α_0, is average electricity expenditures of the zero group (the group for which the value of the independent variable is zero, which here is homes not built in the 1980s) and the estimate of the coefficient on YB1980s, α_1, is the difference in average expenditures between the zero group and the one group (the group for which the value of the independent variable is one, which here is homes built in the 1980s). Adding together the estimates of α_0 and α_1 gives us the average expenditures for homes built in the 1980s. Thus, average electricity expenditures for homes built in the 1980s are: 1,109 + 20 = \$1,129. Average expenditures for homes not built in the 1980s are \$1,109.

Table 2.1 Recoding the variable BUILTYR into multiple binary variables

BUILTYR	*YB1960s*	*YB1970s*	*YB1980s*	*YB1990*	*YB1995*	*YB1999*
1 (1999)	0	0	0	0	0	1
2 (1995–98)	0	0	0	0	1	0
3 (1990–94)	0	0	0	1	0	0
4 (1980–89)	0	0	1	0	0	0
5 (1970–79)	0	1	0	0	0	0
6 (1960–69)	1	0	0	0	0	0

Notes This table illustrates how year built (YB) binary variables would be constructed for six homes, each with a different BUILTYR value

While this model is a starting place, it doesn't make a lot of sense to compare homes built in the 1980s to homes built both before and afterward. Luckily, we can easily modify this model by adding multiple binary variables. Table 2.1 illustrates how to create multiple binary variables based on the variable BUILTYR. This table shows the values of six binary variables, which have names like YB1980s (and which stands for "years built 1980s"), one for each possible value of BUILTYR. For example, if a household has a value of BUILTYR equal to 6, it means the home was constructed in the 1960s, thus the value of YB1960s is equal to 1, and the value of all the other "years built" binary variables is equal to zero.

With the binary variables we constructed from BUILTYR in hand, we can now estimate the following multivariate regression model:

$$COSTELEC_i = \pi_0 + \pi_1 YB1960s_i + \pi_2 YB1970s_i + \pi_3 YB1980s_i + \pi_4 YB1990_i + \pi_5 YB1995_i + u_i.$$

Here, there are five variables on the right-hand side, all of which are binary variables. Notice I did not include all six of the binary variables we constructed. I excluded YB1999. When using multiple binary variables based on an *ordinal* or *categorical* variable, we always need to omit one to prevent an econometric problem called *perfect multicolinearity*. The omitted category then becomes the reference category, what I referred to above as "the zero group." The estimates of this multivariate regression equation appear below:

$$COSTELEC_i = 1{,}056 + 26.6(YB1960s_i) + 89.6(YB1970s_i) + 73(YB1980s_i) + 67.2(YB1990_i) + 11.4(YB1995_i).$$

Because the reference category is homes built between 1999 and 2000, this is the reference group (or zero group). The estimated constant term β_0 is the average electricity expenditures for homes built in this period. All of the other coefficients are differences in means. For example, here we find the coefficient on YB1980s is 73. This means average electricity expenditures in homes built in the 1980s are 73 dollars higher than homes built in 1999–2000. And, we can calculate the average value of expenditures for homes built in the 1980s by adding together the constant term and coefficient on YB1980s (β_0 and β_3: $1{,}056 + 73 = 1130$. This is the same value we found above in the bivariate regression. Here however, the estimates from the multivariate regression can be used to calculate the average electricity expenditures for homes built in all six periods. Figure 2.1 plots average electricity consumption in the year 2000 for single-family homes built in California between 1960 and 2000, by period of construction, using the estimated equation.

Average electricity expenditures are slightly lower in homes built right after the 1970s, as shown in Fig. 2.1. Can we conclude from this that the energy codes California adopted in 1978 were modestly effective? No, because a simple comparison of average electricity expenditures can easily mask the causal effect of energy codes. Despite the fact that these estimates come from a multivariate regression model, it is not a regression control model, and the estimates should be interpreted as descriptive statistics. We can't interpret them as causal effects of energy codes on energy use because of selection bias. As one example, richer households will typically select to live in newer homes, and use more electricity because they are richer. This could make newer homes appear to be less energy efficient than they actually are. Another reason a comparison of these simple averages masks the causal effect of energy codes is that homes have been getting larger over time. Figure 2.2 shows that average home size rises from just over six rooms per home built in the 1960s to nearly seven rooms for homes built in the 1990s.

Figure 2.1 shows average electricity consumption by home vintage, but this doesn't tell us much about the effectiveness of energy codes, and Fig. 2.2 illustrates one reason why. A more informative comparison would be to compare average energy consumption across homes with the same number of rooms. We could even try calculating average electricity expenditures by construction periods using only homes that have the same number of rooms and with occupants that have the same incomes. However, although the ACS is a massive survey, there are few households in the survey

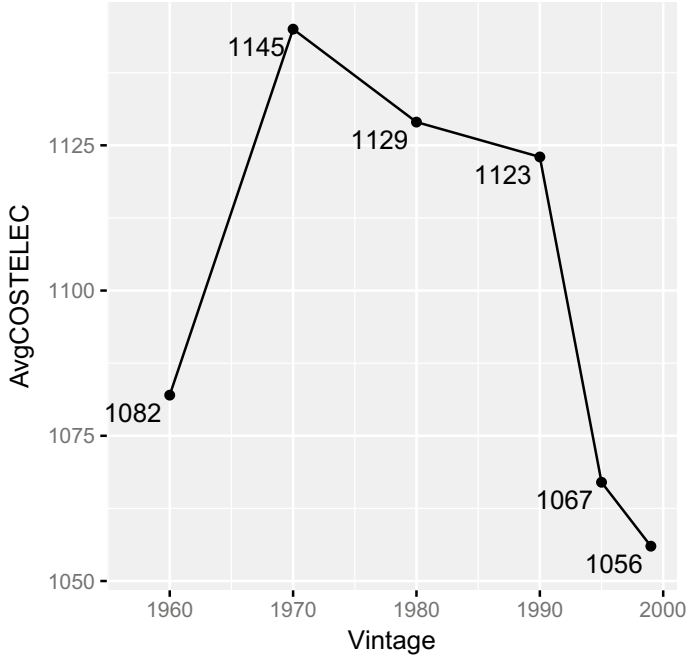


Fig. 2.1 Average ELECOST by homes of different construction eras

that have the exact same number of rooms and income. Thus, in situations like these, researchers turn to the statistical technique of *regression control*.

Regression control includes control variables, like ROOMS and INCOME, in multivariate regression models, so that whatever effect they have on the dependent variable isn't attributed to the main independent variables of interest. Of course, there are numerous other factors beyond home size and occupant income that could influence both a household's decision to live in a more efficient home, and their energy expenditures; in other words, that could bias our estimates. To account for these considerations, Costa and Kahn estimate the more elaborate model shown below:

$$
\begin{aligned}
lnCOSTELEC_{ij} &= \beta_0 + \beta_1 YB1960s_i + \beta_2 YB1970s_i + \beta_3 YB1980s_i \\
&+ \beta_4 YB1990_i + \beta_5 YB1995_i + \beta_6 ELEHEAT_i \\
&+ \beta_7 lnINCOME_i + \beta_8 SEI_i + \beta_9 WHITE_i \\
&+ \beta_{10} ROOMS_i + \beta_{11} HHSIZE_i + \beta_{12} AGE_i \\
&+ \sum_{j=2}^{265} \gamma_j PUMA_j + \varepsilon_i.
\end{aligned}
$$

The goal of estimating this model is that the estimates of the $\beta_0 - \beta_5$ coefficients will be less-biased measures of period-of-construction effects than the $\pi_0 - \pi_5$ coefficients. This model has a lot more variables on the

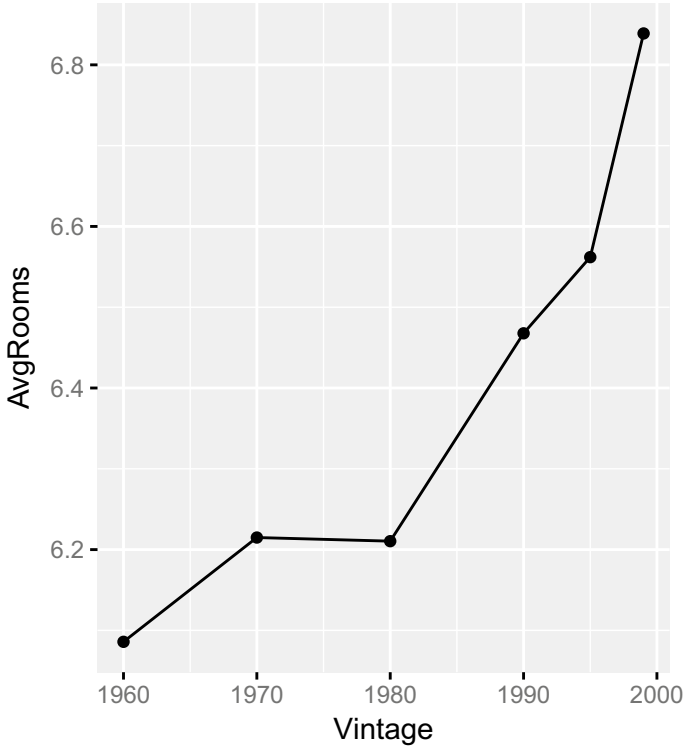


Fig. 2.2 Average number of rooms in single-family homes by construction period

right-hand side compared to the model without control variables. For twelve of the variables shown on the right-hand side of this model, we will estimate a β coefficient on each, plus one more for the constant term, β_0, pronounced "beta zero." (There are also lots of γ coefficients but let's ignore these for now.) In the regression control model, AGE_i is the age of the head of the household, $HHSIZE_i$ is the number of people who live in the household, $ROOMS_i$ is the number of rooms in the home, and $WHITE_i$ is a binary variable equal to one if the head of household is white, and zero if they report any other race.

The variable SEI_i, which stands for "socioeconomic index" is a "composite" variable constructed based on the industry of employment of the household head.[5] The variable $lnHHINCOME_i$ is the natural logarithm of household income, and $ELEHEAT_i$ is another binary variable, equal to one if the household's primary heating fuel is electricity (the most common fuel sources in U.S. households are natural gas and electricity; coal and wood are primary heating fuels for a small minority of Americans).

In addition to these 12 variables, Costa and Kahn's model also included binary variables for all but one of the 265 PUMAs in California (they had to omit one PUMA binary to avoid perfect multicollinearity, just as we saw above when we omitted one of the period of construction dummies).[6] When a multiple regression model includes binary variables for the geographic region the respondents live in, it is called a *fixed effect* model. Here, these PUMA-level variables are called "PUMA fixed effects" or "geographic fixed effects" in the jargon of econometrics.

Having 276 independent variables (the 12 shown in the equation above, plus the 264 PUMA dummies) sounds complicated, but in terms of actually doing it, it takes about five seconds to add them to the code. So it's actually very simple to do, and the computer does the hard work. Including geographic fixed effects is also a very powerful way of controlling for variables

[5] According to IPUMS variable description, "SEI is a constructed measure that assigns a Duncan Socioeconomic Index (SEI) score to each occupation using the 1950 occupational classification scheme available in the OCC1950 variable. The SEI is a measure of occupational status based upon the income level and educational attainment associated with each occupation in 1950. User caution: There is significant debate about the usefulness of composite measures of occupational standing."

[6] As a result, the constant term in the model presented in this section represents a home built in the omitted construction decade (the 1960s), in the omitted PUMA which was "PUMA 060101", an area in the northern portion of Alameda County, which includes the city of Berkeley.

that are hard to measure. For example, the climate in California varies considerably from place to place, and instead of trying to control for climate by measuring it directly, say by determining the average July temperature for each PUMA,[7] we can simply include PUMA fixed effects because all households within the PUMA experience similar climates. Another example is electricity prices. Although Costa and Kahn did not include a measure of the electricity rates households were currently paying, which varies from place to place due to the fact that there are several electricity providers with differing rate structures, these can be controlled for with geographic fixed effects.[8] Suffice it to say that by including geographic fixed effects, we can control for a lot of factors that influence COSTELEC and that are correlated with a home's period of construction, which allows the regression control strategy to better estimate the causal effects we are after.

There is one more difference between the full model Costa and Kahn estimated and the uncontrolled versions I presented above. The dependent variable here is not COSTELEC as it was in the multivariate model without controls, but rather is the natural logarithm of it. The use of a *logged variable* as the dependent variable allows regression estimates to be interpreted as a percentage change.[9] In particular, we can interpret the coefficients on the period of construction dummies as percentage differences from the omitted category (which recall is homes constructed after 1998). And, when both the dependent and independent variable have been logged, coefficients can be interpreted as elasticities. Question 2b at the end of this chapter contains an exercise on interpreting log and other nonlinear models.

[7] This and other PUMA-level climate measures are available from IPUMS-Terra.

[8] Costa and Kahn did include a control variable for the historical price of electricity in the period in which the home was built. I discuss this version of their model in Holian (2020). I don't discuss the version of their model with the historical electricity price variable in this chapter, and instead I focus on their general approach to estimating cohort effects.

[9] The dependent variable $lnCOSTELEC_i$ is the natural logarithm of household i's annual electricity expenditures. Transforming a variable from dollars to the natural logarithm of dollars seems complicated when encountering regression models for the first time, but in practice it is very easy. You probably encountered logarithms in high school algebra. The natural log is log base e where e is about 2.718. The natural log of 1 is zero (raising 2.718 to the power 0 makes it equal 1) and the natural log of 2.718 is 1. In data analysis, one simple line of code transforms all values of COSTELEC to natural logs. Be aware that the log function is not defined for negative values.

Costa and Kahn report the full results from their published article in Table 5 of their working paper (Costa and Kahn 2010). The results I find using their data are shown below:

$$\begin{aligned} lnCOSTELEC_i = {} & 4.396 + 0.137(YB1960s_i) + 0.154(YB1970s_i) \\ & + 0.126(YB1980s_i) + 0.080(YB1990_i) + 0.030(YB1995_i) \\ & + 0.219(ELEHEAT_i) + 0.116(INCOME_i) + 0.066(SEIi) \\ & + 0.083(WHITE_i) + 0.090(ROOMS_i) + 0.056(HHSIZE_i) + 0.005(AGE_i). \end{aligned}$$

These results are identical to those reported by CK.[10] Note none of the estimated γ coefficients are reported, although the constant term captures the fixed effect of PUMA 060101 (the excluded category here). It is conventional to not report the values of geographic fixed effects when they are used as control variables, although sometimes the constant is reported, as it is here.

The fact that the dependent variable has been logged means we can look at the coefficients on the built year dummies and directly interpret them as percentage changes, as follows. The coefficient on YB1960s is 0.137. This means that electricity expenditures are 13.7% higher in homes built in the 1960s, compared to homes built in 1999. Electricity expenditures in homes built in the 1970s are 15.4% higher, and 12.6% higher in the 1980s. The fact that expenditures are lower in the 1980s than the 1970s is better evidence that building codes were effective, but, notice the difference is very small. The difference between the coefficients ($15.4 - 12.6 = 2.8$) means homes built in the 1980s use 2.8 percentage points less electricity. This does not seem like the big difference we'd find if California's building energy codes were very effective.

Another reason the results indicate the energy codes were not very effective is that there is substantial variance in the estimates. As I've mentioned, I don't focus much on statistical inference in this book. But, it turns out there are no statistically significant differences between electricity expenditures of homes built in the 1960s, 1970s, and 1980s. We see this in Fig. 2.3, which shows the "regression controlled" construction period effects. This is a

[10] If you compare these results with their published tables, you'll find the constant terms differ. I omitted PUMA 060101 (Berkeley) while CK selected a different PUMA as the reference category.

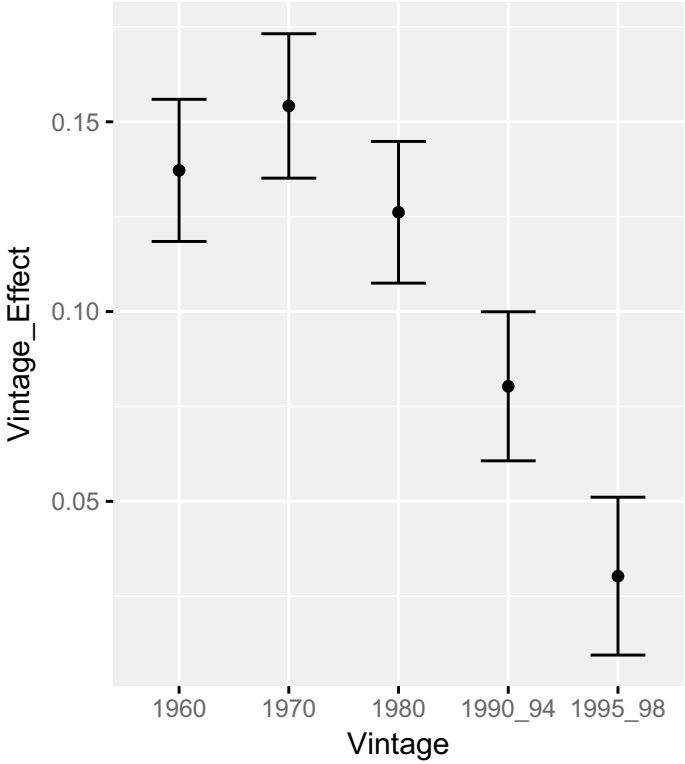


Fig. 2.3 Electricity expenditure "Vintage Effect" in California single-family homes by construction period. The points plot the β_1, through β_5 regression coefficient estimates, and bars show one standard error of the estimated point

plot of the estimated coefficients on the construction era dummies from the equation above, but it also includes the standard error estimates.[11] These error bars overlap, which means the point estimates are not statistically different.

[11] Standard errors are estimates of the variance of the regression coefficients we expect to find in repeated samples.

The challenge with all studies that use regression control to estimate causal effects is knowing whether we have controlled for enough other factors. Costa and Kahn did not, for example, control for how long the household has lived in their home. Failure to control for this could matter, because households tend to add energy-using appliances to the home over time, a point made by Arik Levinson in his 2016 *AER* article, "How Much Energy Do Building Energy Codes Save? Evidence from California Houses." We may find that newer homes use less electricity merely because the residents have not yet installed appliances.

In addition, although Costa and Kahn control for occupational prestige (through the SEI variable) they did not control for the level of education of the occupants. If higher levels of education are associated with greater environmentalist attitudes, then more educated households may decide to live in more energy efficient homes. If these educated, environmentalist households also practice so-called "voluntary restraint" (that is, being careful to limit their energy usage) we may attribute that to the effect of building codes when it is not. While it's possible to critique the model for these omissions, CK did control for many possible sources of bias, thanks to the wealth of variables available in the 2000 decennial Census long form (all of the same variables are also available in the ACS) and the power of fixed effects for small geographic areas (the PUMA fixed effects).

Replications, Extensions, and Original Research

Economists, sociologists, and other social scientists do not agree on the definition of a *replication*. Recent work by Clemens (2017) distinguishes between replication and related studies, and my discussion here draws from this work, but if you compare my definitions to his you will find some differences. Ultimately there are too many nuances to allow for simple definitions that work for all settings. That said, the general idea can be described through some examples.

On one end, if we estimate the same model, with the same sample of data used by the authors of the original study, it would be called a *verification*. This is one type of replication. At the other end, if we estimate a different model, on a different sample from a different population, it is not a replication at all, but rather just *original research*. This is true even if the model is similar (and the data source may even be identical) to one used in a previous study. Related to these polar cases are reproduction, reanalysis, and extension.

An example of a verification would be if you download the data and code files for the Costa and Kahn (2011) study from the AER webpage, and then ran the analysis file. (I describe how to do this in Review Question 1 at the end of this chapter.) This is essentially what we did in the previous section, when I presented the results from the regression control model. There I used the data they supplied, and estimated the exact same model.

In the previous section, I also estimated different versions of the CK model. In particular, I distilled the model down to its most basic component by presenting the bivariate regression model with a single binary variable. To this I added the full set of construction period fixed effects, before eventually presenting a verification of their results. Clemens (2017) would refer to the different versions as a *reanalysis* because the models are different, even though the sample is the same. This is not a replication because we wouldn't expect the results from different models to be the same as those reported in the original study.

Thus in the previous section, we saw both reanalysis and verification of Costa and Kahn (2011). In an article I published in 2020 in a journal called *Economics Letters*[12] I estimated a model that was essentially identical to the model from Costa and Kahn (2011), with the exact same variables and estimation subsample (California homeowners in homes built between 1960 and 2000). While Costa and Kahn (2011) used long-form data from the decennial Census, my study used ACS data from the years 2012 to 2017. (Review Question 3 at the end of this chapter guides the reader through interpreting my reproduction.) This was a *reproduction*, not a verification, because while the population was the same, the sample was different.[13]

The Holian (2020) study also illustrates original research. After presenting a reproduction and extension using the California sample, I then presented results from a "difference-in-difference" (or D-in-D) model that used the full, nationwide sample. I'll hold off on discussing the D-in-D model and results until the next chapter (Review Question 2 at the end of Chapter 3 examines these results). The next three chapters discuss D-in-D in detail; at this point, I'll just highlight that not only was the population under study different (I studied all U.S. homeowners, not just California

[12] Like *AER Papers and Proceedings*, articles in *Economics Letters* are also short and therefore good examples for beginning students.

[13] There are grey areas in all these definitions; here, an open question is whether the sample of data I used is really of the same population. It may not be, because the homes built in 2000 are no longer brand new. If the population is different, then what I did in Holian (2020) is not a reproduction but an extension, using the Clemens (2017) taxonomy.

Table 2.2 Defining replication

		Same population?	
		Yes	No
Same model?	Yes	Replication[a]	Extension
	No	Reanalysis	Original Research

[a] Includes both verifications and reproductions

homeowners), the model is different as well, and therefore this is an example of original research, not an extension and not a replication. This is true, even though I obtained data from IPUMS, the same sources as Costa and Kahn (2011), and some of the research questions were the same.

The discussion so far is summarized in Table 2.2.

Although categorization is not always as black and white as Table 2.2 suggests, it helps make sense of the discussion so far. I have two final examples. In some of my research in progress, I estimate a model that is essentially identical to the one from Costa and Kahn (2011) shown in the previous section, but that uses 2012–2017 ACS data from Florida homeowners. Florida passed its first building energy codes in the late 1970s and so is similar to California in this regard, but Florida is a different population than California. Therefore this would be considered an *extension* in the taxonomy of Clemens (2017). Another example that would be considered an extension would be if I estimate the same model as Costa and Kahn (2011) but using just customers in San Diego's electricity district. The San Diego population is different from the California population. This would not be a reanalysis, but an extension that would be possible to do even with the author-supplied data (i.e. it would not require obtaining the original raw data, though doing so is often a best practice).

This discussion is nuanced, but the main point I want to emphasize is that there is a path researchers can follow from replication to original research. Mapping this path for the reader is a goal of this book. If you can carry out a verification of a study that uses ACS data, it's not that hard to obtain the full, raw data from IPUMS and carry out an extension of the study, or original research based on it. Whether you then "sell" your research as a replication or original research is up to you.

A few final words on the importance of moving beyond author-supplied data. It may seem like it doesn't matter much whether I obtain the data from IPUMS myself, or just use the data supplied by the authors, but the

difference is huge. Sometimes authors inadvertently modify data after they obtain it and before they share it. From time to time IPUMS fixes errors it notices; the data the original authors used may have contained errors that have since been corrected.

More importantly, because I obtained the original, raw data myself, I am now in a position to do more extensions and original research than I would be able to do if I just used the data shared by Costa and Kahn (2011). For example I can estimate the same model they did but for Florida, and use natural gas expenditures rather than electricity expenditures as a dependent variable. Both electricity and natural gas use contribute to household carbon footprints. Whether we classify this research as an extension or original research ultimately may not really matter, because in both cases we are creating new knowledge.

The findings from the CK study regarding energy codes, as well as in my analysis of the topic using the ACS (see also Holian 2021), suggest that energy efficiency regulations were not as effective in reducing electricity consumption as predicted when they were adopted. This doesn't mean we should scrap them or stop trying to strengthen them, or that future energy codes couldn't be more effective. We will return to the question of whether building energy codes are wise policies in the concluding chapter of this book on Cost-Benefit Analysis.

The main takeaway from this discussion of regression control is that there are often factors that are correlated with our main independent variable of interest that also affect the dependent variable. If these factors are not included in our analysis, they will bias our estimate of the coefficient on our key variable of interest. If we can measure these factors directly, multivariate regression offers a convenient way to include them in the analysis–we just add them to the right-hand side of the regression equation. If we can't measure the factors directly, we can often still control for them through the use of geographic or other fixed effects, as we saw in this chapter.

The next three chapters focus on a different econometric technique for estimating causal effects, called difference-in-differences (D-in-D). While this technique is the focus of the case studies in Chapters 3 through 5, these chapters also discuss some studies that use regression control techniques. For example, in Chapter 3 we discuss an article (Winters 2017) that uses regression control to study migration between labor markets of college graduates within the U.S. In Chapter 5, we discuss a study of the effect of marriage and children on a woman's labor-market outcomes (Marshall and Flaig 2014). In fact, an entire section of Chapter 5, titled, "Marriage,

Self-Employment and the Gig Economy" focuses on further details of the regression control approach. Regression control is a widely used technique, and we'll return to it at various points throughout this book.

The topic of the next chapter is migration. The focus is on the contemporary period, where migration from Central America is the immigration debate of the day. From 2016 to 2019 there was a controversy surrounding the asking of a citizenship question on the 2020 decennial Census. The main purpose of the decennial Census, which does not ask a citizenship question, is to reapportion the U.S. House of Representatives. The ACS has asked a citizenship question every year since 2000. We begin the next chapter with stories about migration to the USA, examples of how the citizenship question is used in research designs, and some statistics that shed light on the effects of immigration policy on labor-market outcomes.

Key Terms

Regression control	Ordinal variable	Perfect multicollinearity
Fixed effect	Logged variable	Replication
Verification	Original research	Reanalysis
Reproduction	Extension	

Questions for Review

Questions 2–4 below are empirical exercises. A reader may wish to read Appendix A before attempting them.

1. Explain whether the regression coefficient estimate of π_3 from this chapter is a good estimate of a home's energy efficiency, or it suffers from OVB. List an omitted variable that satisfies the two conditions for OVB. Explain how the regression control technique is used to estimate a β_3 coefficient which is less biased than π_3.
2. This question guides you through verifying the two regression models in Table 1 of the Costa and Kahn (2011) study, using data supplied by the authors, and carrying out some reanalysis and extensions of it. First, download the data at https://www.openicpsr.org/openicpsr/project/112425. There are 11 files. The two data files you need for

this question are: `Table 1data.dta` and `price57.dta` (these are Stata data files). Create a directory folder and place these files in it. If you have Stata, try running the code in a third file `Table 1.do`. To use R instead, download the file `script5.R` from this book's webpage:

(a) Run the analysis file. Compare and contrast the results you get with the published results in terms of sample size and magnitude of estimated coefficients.
(b) Modify the model by making the dependent variable linear (use COSTELEC rather than logCOSTELEC). Do the same for income. Interpret this model.
(c) Modify the model by changing the reference category (omit YB1960 rather than YB1999). Carry out and interpret a hypothesis test for YB1980.
(d) Can you think of any extensions you can carry out with the author-supplied data? Hint: the variable "fips" in the merged data identifies the county in which the respondent lives (e.g. 37 is the FIPS code for Los Angeles County).

3. This question guides you through verifying the first regression model in Table 1 of the Costa and Kahn (2011) study, using data you download from IPUMS-USA yourself. Access the IPUMS-USA website at https://usa.ipums.org and register for a free account. Navigate to "Get Data" and click "Add Samples" and select the 2000 5% sample. Next, select "Add Variables," and add 15 variables: AGE, BUILTYR, COSTELEC, COSTGAS, FUELHEAT, HHINCOME, NUMPREC, OWNERSHP, PUMA, RACE, RELATE, ROOMS, SEI, STATEFIP, UNITSSTR. To download the file, select "View Cart" and then "Create Data Extract". Make sure to select CSV under Data Format. In addition to the 15 variables you selected, you'll get 9 additional preselected variables, and 3 detailed versions. Finally, download the file `script6.R` from this book's webpage.

(a) Run the analysis file. Compare and contrast the results you get with the published results in terms of sample size and estimated coefficients.
(b) Carry out some reanalysis of the data. Some suggestions: Do PUMA effects matter? Do clustered standard errors matter? Does weighting matter?
(c) Change the estimation subsample from California to Florida homeowners. Is this an extension or original research?
(d) Change the dependent variable from logCOSTELEC to logCOSTGAS. Is this an extension or original research?

4. These questions are about the reproduction of the CK model carried out in Holian (2020). Download the file `script7.R` and find links to the published or working paper versions of the article on this book's webpage. This R script can be run with the master ACS file described in Appendix A. Verify you get the results in the published version of the study for the California regressions. Discuss differences between the CK model and the models in Holian (2020) in terms of: what is the excluded category, which control variables are used, and what the results say about how effective California's adoption of building codes were in reducing electricity consumption.

References

Clemens, Michael A. "The meaning of failed replications: A review and proposal." Journal of Economic Surveys 31, no. 1 (2017): 326–342.

Costa, Dora L., and Matthew E. Kahn. Why has California's residential electricity consumption been so flat since the 1980s? A microeconometric approach. No. w15978. National Bureau of Economic Research, 2010.

Costa, Dora L., and Matthew E. Kahn. "Electricity consumption and durable housing: Understanding cohort effects." American Economic Review: Papers & Proceedings 101, no. 3 (2011): 88–92.

Elliott, Catherine S. "A May American Economic Review papers seminar and an analytic project for advanced undergraduates." Journal of Economic Education 35, no. 3 (2004): 232.

Holian, Matthew J. "The impact of building energy codes on household electricity expenditures." Economics Letters 186 (2020): 108841.

Holian, Matthew J. "Corrigendum to 'The impact of building energy codes on household electricity expenditures' [Econom. Lett. 186 (2020) 108841]." *Economics Letters* 200 (2021): 109738.

Levinson, Arik. "How much energy do building energy codes save? Evidence from California houses." American Economic Review 106, no. 10 (2016): 2867–2894.

Marshall, M.I., and A. Flaig. "Marriage, children, and self-employment earnings: An analysis of self-employed women in the US." Journal of Family and Economic Issues 35, no. 3 (2014): 313–322.

Nordhaus, William D. The climate casino: Risk, uncertainty, and economics for a warming world. Yale University Press, 2013.

Winters, John V. "Do earnings by college major affect graduate migration?" The Annals of Regional Science 59, no. 3 (2017): 629–649.

PART III

Difference-in-Differences

Learning Goals for Part III

1. Discuss how the difference-in-differences (D-in-D) technique relies on natural experiments to determine which individuals receive treatment of a policy or event, in a way that is as-if treatment was assigned by an experimenter.
2. Describe how the basic D-in-D model was used in the case studies in Chapter 2–4 to study the effect of: immigration policy on employment among Salvadoran immigrants, the Affordable Care Act on entrepreneurship, and the Great Recession on fertility.
3. Discuss how a multivariate regression model with an interaction term can be used to estimate the four group means required to produce the basic D-in-D estimate.
4. Compare and contrast: (i.) the basic D-in-D model, (ii.) the basic D-in-D model with control variables, (iii.) the fixed effect D-in-D model, and (iv.) the two-way fixed effect estimator (TWFE).
5. Discuss a limitation of D-in-D approaches to estimating causal effects.

CHAPTER 3

Searching for Higher Ground: Migration and Quality of Life

Migration is a source of contemporary social controversy in the USA and around the world. Britain's 2016 vote to withdraw from the European Union, dubbed, "Brexit," was driven by anti-globalist and anti-immigrant sentiment, and in the same year in the USA, immigration from Central America sparked a political backlash that helped elect Donald Trump as president. The American Community Survey (ACS) provides a valuable window into migration-related issues, including illegal immigration, migration across states, and settlement within cities. The ACS asks respondents about their place of birth, where they lived last year and, a highly sensitive topic for some respondents, their citizenship status. Among immigrants, the ACS also asks how long they've been living in the USA. This chapter describes migration research that analyzes the ACS data, using the core econometric methods of *regression control* and *difference-in-differences* (D-in-D).

The case study in this chapter, like the case studies in Chapters 4 and 5, illustrate the D-in-D technique for causal inference. Both regression control (which was the focus of the last chapter) and D-in-D are used by researchers to estimate unbiased (or less biased) causal effects, but they differ in that regression control requires obtaining data on control variables, while D-in-D requires identifying a natural experiment that separates people into control and treatment groups. The D-in-D technique is a widely used tool among researchers analyzing the ACS data, which is why Part III of this

M. J. Holian, *Data and the American Dream*,
https://doi.org/10.1007/978-3-030-64262-4_3

book has more chapters than all other parts. This chapter illustrates the technique by way of a case study of immigration policy.[1]

Many discussions of immigration we hear in our everyday lives or in the media often seem to be of the general, "immigration is good" or "immigration is bad" variety. The studies we discuss in this chapter, in contrast, measure impacts of specific pieces of immigration policy, and are therefore better suited for study with Cost–Benefit Analysis, which I discuss here and in the concluding chapter.

Several of the studies I discuss below relate in one way or another to the so-called "back to the cities" phenomenon. While the urban landscape in the latter half of the twentieth century was characterized by an exodus of households from cities to the suburbs, we now seem to be witnessing a revival of American cities. Many downtown areas are experiencing a building boom, fueled by high housing prices. While there are positive aspects of this urban renaissance, residents who never left cities now have to compete with new arrivals for housing. The name given to the process of higher income households moving into formally lower income neighborhoods is "gentrification." Both illegal immigration and gentrification are situations where one person's striving for a better life can conflict with someone else's ambitions.

Before turning to the discussion of research that uses the ACS to study migration, I'll share a migration story from a woman I'll call Linda whose family came from Honduras. My goal in including stories like these in this book is to keep the reader focused on the humans represented in the data. One of the studies we'll discuss in the next section has to do with an immigration policy known as Temporary Protected Status (TPS) which is a temporary authorization given to migrants from certain countries, usually following a catastrophic natural disaster or violent conflict. Linda is from Honduras, a country which has had a TPS designation since 2001, and I asked Linda if she knew anyone with TPS.

She did. It turns out she has two nieces from Honduras who arrived in the USA a few months apart. Because of the timing of the separate arrivals

[1] One of the reasons the D-in-D techniques is so often used to study the ACS is because D-in-D typically requires data that has a time-dimension. The ACS fulfills this requirement because it is a *repeated cross-section*, meaning, it's a survey (cross-section) that takes place every year (repeated). Survey year is thus one obvious ACS variable that has a time dimension, in a sample with multiple years of ACS data. Other sources of time variation can be found within a survey year, including questions that reveal an individual's birth cohort or the year of construction of a home, as we saw in the last chapter.

of these sisters some 20 years ago, today one qualifies for TPS and the other does not. I asked Linda if she thinks their different immigration status has had an impact on their lives. Her eyes get big as she tells me the difference between them couldn't be greater.

The niece who does qualify for TPS has a salaried job with full benefits. The niece who does not qualify has been living for the last two decades as an unauthorized (undocumented) immigrant. This has severely hampered her ability to find work. Most recently, she worked for $12 an hour, which is less than the minimum wage in San Francisco. Although it is illegal for an employer to pay less than minimum wage, her unauthorized status means her employers often ignore minimum wage laws.

Poor job opportunities are not the only challenge the unauthorized niece faces. Her husband, also an unauthorized immigrant, was deported after being falsely accused of a crime. Her daughter was placed in foster care for a period, and today lives with the fear that her mother could be deported any day. Meanwhile, her sister who immigrated just a few months apart from her, is living the American dream.

Immigration is a key part of the story of many Americans, and this was captured-well by the singer-songwriter Jimbo Scott. He tells three immigration stories in his song, "Live Free": that of an Irish women arriving at Ellis Island fleeing the 1845 potato famine, a Chinese man held at Angel Island in the San Francisco Bay under the 1882 Chinese Exclusion Act, and a Central American women at the southern border in the Trump era. "They come from all directions…" he sings, and, "…these are the American Dream."

Many Americans are decedents of immigrants. But America is not just a nation of immigrants. It is also a nation of the descendants of African slaves that were brought here against their will, and indigenous people who were here before our political boundaries were drawn. Martin Luther King Jr. once wrote, "We all came in on different ships, but we're all in the same boat now." There are no easy answers to the immigration policy questions of our time.

Border Walls, Immigration Enforcement, and Work Permits

Controversy over illegal immigration has intensified in recent years, but is nothing altogether new. In the 2016 U.S. presidential election, Donald Trump campaigned on building a wall along the southern border which

proved extremely controversial, but it was often forgotten that The Secure Fence Act was ratified by Congress and signed into law by President George W. Bush in 2006. This Act added 548 miles of wall to the U.S.–Mexico border.[2]

A number of studies use the ACS to measure the impact of immigration policies. Kostandini et al. (2013), in a study titled, "The impact of immigration enforcement on the US farming sector," note the U.S. farming sector heavily depends on seasonal workers and undocumented immigrants. These authors find that, when some local jurisdictions (such as counties and states) began enforcing federal immigration laws, under the so-called "287(g)" program, passed in 1996 with implementation starting after 2002, the presence of non-citizen immigrants in these communities decreased. The basic supply and demand model for wages predicts that, when the supply of workers in an industry falls, wages increase for the workers who remain, and Kostandini et al. (2013) present evidence consistent with this for the agricultural sector. Although most studies I discuss here focus on the impacts of immigration policies on immigrants, it is relevant to separately highlight the impacts of immigration policies and their enforcement on citizens, because they are the ones who vote and therefore drive policy changes in this area.

Amuedo-Dorantes et al. (2018) study a broader range of enforcement programs using the ACS, in addition to and including the 287(g) programs, in an article titled, "Immigration enforcement and economic resources of children with likely unauthorized parents." They find increased local enforcement of federal immigration policies increased poverty rates among families with parents who are likely unauthorized.[3] Many of the children in these households were born in the USA even though the parents were not, and this suggests a different political economy impact that will take effect over time as these children reach voting age.

[2] A 2019 study by Treb Allen, Melanie Morten and Caue Dobbin, titled, "Border Walls" presents estimates of wage impacts due to the walls constructed as a part of this 2006 legislation. In addition, they use the technique of Cost–Benefit Analysis to assess whether or not the investments in border walls were worthwhile. Although their CBA is not fully comprehensive (for example it ignores impacts such as possible reductions in crimes that would have been committed by illegal immigrants), it is a notable example of the use of reason over symbolism in this contentious area.

[3] While the ACS asks immigrant respondents if they are citizens, the ACS doesn't ask them about their legal status. Thus Amuedo-Dorantes et al. (2018, p. 66) describe how they "...rely on Hispanic ethnicity and lack of citizenship as a good proxy for likely unauthorized status".

A third immigration study to use the ACS examines changes in legal status rather than changes in enforcement. Elira Kuka and her coauthors explore the impact of the Deferred Action for Childhood Arrivals (DACA) program in an article titled, "A Reason to Wait: The Effect of Legal Status on Teen Pregnancy." The authors note (on p. 213) that there are many reasons why an undocumented teenage woman may have a child, including that, "...undocumented women may use birthright citizenship and have children to prevent deportation." Birthright citizenship refers to the constitutional right of all children born in the USA to be citizens, and has been a feature of American policy for over a hundred years. (See Orrenius and Zavodny 2010, p. 108 for an economics view on this issue.)

The DACA program granted temporary authorization to those who qualified, and this could conceivably reduce a teenage woman's incentive to have a baby for several reasons. DACA made staying in school more attractive because school enrollment is one requirement for qualifying for DACA. The authors indeed find that the incidence of teenage pregnancy fell dramatically among Hispanic immigrant non-citizens (some of whom were eligible for DACA) after it was implemented, relative to Hispanic immigrant citizens (who were not affected by DACA).[4]

All three of these studies make use of a technique that is the focus of Part III of this book, which is called the *difference-in-differences* (or D-in-D). Of the three, the Kuka et al. (2019) study would be the best choice for a beginning student to replicate, for a few reasons. First, author-supplied replication files for the study are available on the *AEA Papers and Proceedings* website. Having these files doesn't always make replicating a study straightforward but having them can still be helpful. Second, their model can be estimated using only publicly available data from IPUMS. Third, Kuka et al. (2019) estimate a basic D-in-D model.[5] The Kostandini et al. (2013) and Amuedo-Dorantes et al. (2018) studies, on the other hand, are less ideal for a beginning student to replicate, because authors-provided replication files are unavailable, both use merged data that can be recol-

[4] As in the study by Amuedo-Dorantes et al. (2018), Kuka et al. (2019) do not actually know whether someone in the non-citizen immigrant group is unauthorized, only that many in this group are unauthorized.

[5] Technically, the model they estimate is a basic D-in-D model with control variables, including fixed effects which were discussed in the previous chapter. But, it would be possible to carry out a reanalysis of their model that simply drops these control variables, and what would be left is a basic D-in-D.

lected but is not readily available, and finally (and most importantly) both of these studies estimate a complicated version of the D-in-D model, called a *two-way fixed-effect* (TWFE) estimator.[6]

I now turn to an extended discussion of a fourth study of illegal immigration which uses the ACS, and which is the study I replicated for this chapter. It is titled, "The impact of temporary protected status on immigrants' labor market outcomes," by Pia Orrenius and Madeline Zavodny. This study, also published in the *AEA Papers and Proceedings* journal, estimates a basic D-in-D model. It considers the impact of a policy that grants temporary authorization to undocumented immigrants; this is the Temporary Protected Status (TPS) policy discussed earlier. Recall, TPS is designated for countries experiencing war or other hardship, and it means unauthorized immigrants from those countries that meet specific criteria regarding the timing of their arrival do not face deportation and can work legally until the TPS expires.

The Orrenius and Zavodny (hereafter OZ) study focuses on migrants from El Salvador. Salvadorans are the largest group of immigrants to hold TPS. The OZ study exploits the "natural experiment" that separates Salvadoran migrants into control and treatment groups, based on whether they arrived before or after 2001 and were or were not eligible for TPS. Immigrants from El Salvador were granted TPS in March of 2001. OZ thus study outcomes for migrants from El Salvador who migrated between 1999 and 2000 and qualified for TPS, and those who migrated between 2002 and 2003 and did not qualify. They exclude migrants who arrived in 2001 from their estimation subsample because TPS came into effect in March of that year, making it impossible to correctly determine if individuals arriving in 2001 qualified for TPS or not.

Due to the massive scale of the ACS, there is a large sample of Salvadoran migrants with these entry dates in the sample. The sample size is even

[6] The TWFE estimator allows there to be more than two periods and more than two entities, and also allows the timing of the policy to differ across the entities. The reason I describe the Kuka et al. (2019) model as an example of basic D-in-D is that the DACA policy went into effect at the same time for all individuals impacted by it. In the other two studies, local jurisdictions began enforcing federal immigration law at different times. The main complication with a TWFE model is in interpreting what the estimates mean. In fact, the question of how best to interpret results from a TWFE model is a problem the econometrics literature continues to wrestle with; see Goodman-Bacon (2019). Callaway and Sant'Anna (2020) in section 5 of their paper replicate a study that uses the TWFE and carefully reanalyze it using data, code and program files the authors have made available. In this book, I provide several examples of TWFE papers that could be reanalyzed in this manner.

enough to study men and women separately, by skill level. Let's consider one of their results concerning a group of less educated females (defined as those who did not attend college). The study considered a variety of outcome measures, including wages and work hours. Concerning employment, 62% of Salvadoran noncitizen immigrants that arrived between 1999 and 2000 (many of whom were eligible for TPS) were employed in the 2005–2006 period. OZ use data from these two years of the ACS. However, among the group that came during 2002–2003 and who do not qualify for TPS, only 40.6% were employed in this same period. This is a large difference of 21.4 percentage points (PP). But is this 21.4 PP difference a causal effect of TPS on employment among less-educated female Salvadoran migrants?[7]

One might reason that, because the group that had TPS came earlier, they may have had more time to adjust to life in America and thus to find work. Perhaps the 21.4 PP difference is only partially explained by the TPS policy. To address this possibility, OZ compare the employment rate difference of 21.4 to that of Mexican migrants that arrived during the same periods. Mexican migrants have never been granted TPS, but OZ argue they are similar to Salvadorans in most other respects and thus serve as an adequate "control" group (in the terminology of D-in-D, the control group is "never treated," while the treatment group, here Salvadorans, is treated in one period and not the other).

Among Mexican migrants, the employment rate was 36.1% for those who migrated in the 1999–2000 period, and it was somewhat lower, at 31.6%, among those who migrated more recently in the 2002–2003 period. This is a difference of 4.5 PP. This suggests the 21.4 PP difference we calculated above for less-educated female Salvadoran migrants is an overestimate of the causal effect of the TPS on employment.

The discussion so far has set up all the ingredients necessary to understand what economists call the D-in-D estimator. The first three rows of Table 3.1 below reproduce the statistics discussed above. The last line is new, and I discuss it next.

[7] In the terminology of D-in-D, Salvadorans are the treatment group, and the ones that came in the earlier period were "treated" by the policy. 21.4 is both a difference in means and a difference in proportions. Consider a binary variable equal to one if the person was employed and zero if not. Then the mean of this variable equals the proportion of people that are employed.

Table 3.1 Employment rates among less-educated migrant women

	El Salvador	*Mexico*
Migrated 1999–2000 (TPS eligible)	62.0	36.1
Migrated 2002–2003 (Not TPS eligible)	40.6	31.6
Difference	21.4	4.5
Difference-in-differences	21.4 − 4.5 = 16.9	

In the counterfactual world where nothing happened to the legal status of Salvadoran migrants, we might have expected employment rate differences between early and later arrivals to mirror those of Mexicans, namely, the employment rate would have been 4.5 PP higher among the earlier Salvadoran arrivals. In particular, employment would have been 40.6 + 4.5 = 45.1%. We actually found it to be 62%, so employment among the earlier arrivals was 62 − 45.1 = 16.9 PPs higher than it would have been without TPS. This brings us to the last line of Table 3.1, which presents another way of calculating the D-in-D estimate of 16.9. There, we take the difference between earlier and later Salvadoran employment rates (21.4 PP) and subtract the difference between earlier and later Mexican employment rates (4.5 PP) to again find the D-in-D estimate: 16.9.

The upshot is, a simple difference in means analysis would find TPS accounts for 21.4 PP higher employment, when the figure may really be closer to 16.9 PP. The D-in-D method is sometimes taught as an advanced method, but we see here it is really just the difference between two difference in means. While someone could spend a lifetime studying the nuances of D-in-D, no fancy concepts other than that of averages and subtraction are needed to understand it in its most basic form.

We saw in the last chapter that regression models are often a convenient way of calculating means and differences in means, and this remains true in the case of D-in-D. Earlier we saw that, when there are two groups, a bivariate regression model with a binary independent variable enables us to estimate a difference in means. In the case here, we have four groups, Salvadoran migrants with and without TPS, and Mexican migrants who arrived between 1999–2000 and 2002–2003. We'll need a specific type of multiple regression model to estimate each of the four group means, which is shown below:

$$EMP_{it} = \alpha + \beta ELSALV_i + \gamma EARLY_t + \delta(ELSALV_i \times EARLY_t) + \varepsilon_{it}.$$

Here the dependent variable EMP_i is a binary variable equal to zero if individual i is not employed, and equal to one if they are. The variables

on the right-hand side are also binary variables. $EARLY_i$ is equal to one if individual i came in during 1999–2000 and zero if they came in during the later period of 2002–2003. The variable $ELSALV_i$ is equal to one if individual i is from El Salvador and zero if they are from Mexico. The third variable on the right-hand side is the product of $EARLY_i$ and $ELSALV_i$; the product is also a binary variable; it will be equal to one only if both variables are one, which is true for individuals who came in during 1999–2000 and are from El Salvador, and it is zero otherwise. Because this model features a variable which is the product of two variables, it is often called an *interaction model*.[8]

All basic D-in-D models we'll encounter in this book have the same general form as the interaction model shown above. These models all have a binary variable that indicates the timing of the policy, and they also all have a binary variable indicating which group receives the treatment. Finally, they all have an interaction term, and the coefficient on this interaction term is the key coefficient of interest because it indicates how the policy affected the treatment group (more precisely, the δ coefficient measures the average effect of treatment on the treated).[9]

When we estimate this equation, we find estimates of the coefficients. The estimated equation of this basic D-in-D model is shown below:

$$EMP_{it} = 0.316 + 0.090\,(ELSALV_i) + 0.045\,(EARLY_t) + 0.169\,(ELSALV_i \times EARLY_t).$$

To interpret this model, we can calculate fitted values. Plug in values of zero and one for the variables on the right-hand side to calculate the four employment rates that are shown in Table 3.1. Let's consider first a migrant from Mexico who came in between 2002 and 2003. Because they came in after the TPS period, the value of $EARLY_i$ is zero, and because they are

[8] In the last chapter we saw a multivariate regression model with many independent variables. The interaction model here is just a multivariate regression model with three independent variables on the right-hand side, where one of the independent variables happens to be the product of the first two. The fact that this model has a binary dependent variable on the left-hand side makes it a Linear Probability Model (LPM) which we first encountered in Chapter 1, Review Question 8. With LPM models, we interpret fitted values as predicted probabilities.

[9] In most settings, the policy is in effect after a certain date, not before a date, as is the case in the TPS example. Thus in most D-in-D settings, we'll name the timing variable POST. It is also typical to name the variable that indicates the treatment group as TREAT. Question 3 at end of this chapter discusses an equation with this standard naming convention.

from Mexico, the value of $ELSALV_i$ is also zero. Thus plugging in zeros for all the right-hand side variables, the equation reduces to 0.316. This is interpreted as the probability of employment for a less-educated female Mexican migrant who came in during the 2002–2003 period. It is the same value reported in Table 3.1, but the figure in the table converted the probability of 0.316 into a percentage by multiplying it by 100. Next, let's consider a Mexican migrant who came in during the 1999–2000 period. The value of $EARLY_i$ for them would be 1. However because $ELSALV_i$ would be zero for this individual, the equation reduces to 0.316 + 0.045 = 0.361. This is interpreted as the probability of employment for a less-educated female Mexican migrant who came in during the period when TPS was in effect for Salvadoran immigrants, reported in Table 3.1 as 36.1%. As a check on understanding, readers should verify they can use this equation to calculate the employment rates reported in Table 3.1 for less-educated female Salvadoran migrants who came in before and after the TPS period.

At the beginning of this chapter, I indicated D-in-D requires identifying a natural experiment that separates people into control and treatment groups, and that this differs from regression control, which requires obtaining data on control variables. It turns out, there is a role for control variables in the D-in-D framework. (I hinted at this above in footnote 5.) In their study OZ estimate a *basic D-in-D with control variables*. Why? There are several reasons. One is that, say the composition of migrants from El Salvador changed after 2001. Control variables can be used to limit the bias this introduces, at least from sources that can be measured.[10] OZ include control variables for marital status, educational attainment, age, and fixed effects for the state in which the migrants live, in addition to the core variables in the basic D-in-D model shown above. Formally, their equation can be written as follows:

$$EMP_{ist} = \alpha_c + \beta_c ELSALV_i + \gamma_c EARLY_t + \delta_c (ELSALV_i \times EARLY_t)$$
$$\beta_1 MARRIED_i + \beta_2 DIVWIDSEP_i + \beta_3 LESSHS_i + \sum_{k=Alaska}^{Wyoming} \beta_k STATE_k\, s$$
$$+\beta_4 AGE_i + \beta_5 AGE_i{}^2 + \beta_6 AGE_i{}^3 + \beta_7 AGE_i{}^4 + \varepsilon_{ist},$$

[10] Another reason is adding control variables makes our fitted values closer to the true values observed for respondents, and this leads to smaller standard errors on the estimated coefficients, which in turn makes the coefficients more likely to be statistically significant.

and in Table 3.1 of their article they report the estimated value of the δ_c coefficient. In this equation, the top line variables are all as before, the middle line includes three binary control variables (MARRIED is equal to one if married and zero if not, DIVWIDSEP is equal to one if divorced, widowed, or separated and zero if not, and LESSHS is equal to one for no high school degree and zero if high school diploma) and 50 state fixed effects. The fixed effects are similar to the PUMA-level fixed effects we saw in Chapter 2, but here they are at the state level.[11] The final line of the equation contains a *polynomial* specification of the AGE variable; rather than just include an AGE variable by itself, the authors also include the square of it (and the third and fourth-order terms). Doing this allows for the effect of age to be nonlinear in the model.[12]

The file `scripts8.R` on this book's webpage estimates this equation. Using it I find the estimate of δ_c is 0.173, this is the value OZ reported in Table 3.1 (row 1, column 2). This estimate of 0.173 is very close to the figure of 0.169 that we found above in the basic D-in-D model (without controls). It turns out, the results from the two models are similar, but we wouldn't have known this without estimating the more complicated model. The fact that the results don't change much when we add control variables boosts our confidence that the findings are not sensitive to modeling choices.

As another check on whether their estimates represent causal effects, OZ carry out a *falsification test* using an alternative control group—Guatemalan migrants. The D-in-D technique assumes that absent TPS, the difference in employment between early and later Salvadoran migrants would equal the corresponding difference among early and later Mexi-

[11] If the summation notation confuses you, try replacing the term $+\sum_{k=Alaska}^{Wyoming} \beta_k STATE_{ks}$ with the following term: $\beta_{AK} STATE_{AKi} + \beta_{AZ} STATE_{AZi} + ... + \beta_{WY} STATE_{WYi}$. Both ways of writing it convey an identical meaning, the former is just more compact. There are 51 state entities (including the District of Columbia) so we have 50 binary variables after excluding the reference category. At most one of these binary variables will be equal to one for a respondent, and the rest will be zero.

[12] Polynomial transformations are easy to implement in practice; similar to the log transformation we saw in Chapter 2, it requires first creating a new variable which is the square (or cube) of AGE, and then including this new variable along with AGE in the model. Polynomial models are actually much more flexible than log models for estimating nonlinear effects, but the advantage of log models is there are rules of thumb that make them easy to interpret, whereas interpreting polynomial models takes a little more care. Luckily, we don't usually have to worry about carefully interpreting the coefficients on control variables, because they are only included to give us a better estimate of the coefficient on our main variable of interest.

can migrants. They argue, "Guatemala is similar to El Salvador in a number of respects, including a weak economy and widespread violence, but Guatemalan migrants have never received TPS. Like Mexicans and Salvadorans, most Guatemalan migrants are unauthorized immigrants" (pp. 577–578). They do not find any significant differences between Guatemalan migrants who came in before and after the treatment period, relative to Mexican migrants who came at the same times, and this suggests (but does not prove) that the figures for Salvadoran migrants are the result of TPS.

The findings from the OZ study suggest that TPS dramatically increased employment for less-educated Salvadoran female migrants by 17 PP. Labor-market outcomes have real consequences for these workers and for their families. To more fully appreciate this, consider again the story of Linda and her nieces from the first section of this chapter. The real differences in the quality of life between Salvadoran migrants with and without TPS could easily be much greater than is suggested by the 17 PP figure, which may seem like a moderate impact to some readers. But interpretation of statistics is part of the art of econometrics, and given what we know about the challenge of living in the USA as an unauthorized immigrant, from stories and anecdotes, I would suggest we interpret this figure as highlighting dramatic differences that are caused by the policy.

Does this figure suggest what we should do in this area of immigration policy? The figures represent one impact of the policy, but they do not say anything about other costs and benefits of the policy, let alone who has standing. Immigration policy has far ranging impacts that are not well understood. Earlier we encountered the Allen et al. (2019) study, which illustrates how economists are grappling with conducting Cost–Benefit Analysis in the complex policy area of immigration.

Back to the Cities?

The fraction of Americans living in suburbs had been mostly increasing since the invention of the automobile. In the year 2000, 50% of Americans lived in suburban areas of metropolitan areas, while only 30% lived in the central cities of metro areas (the other 20% lived in rural areas, outside the metros). Between 2000 and 2010, the average annual population growth rate was 1.38% in suburbs and only 0.42% in central cities. In 2010, however, things began to change. Each year, from 2010 to 2015, central cities

grew faster than suburbs, for the first time in decades, causing some to ask, "Will This Be the Decade of Big City Growth?"[13]

Although the period from 2010 to 2015 saw faster growth in central cities than suburbs, from 2016 to 2019, suburbs again grew faster than central cities in each year. However, there remains a great deal of nuance in describing dynamics of urban America. That something different is happening in our cities is undeniable to anyone who has walked through them, but good, descriptive measurement of this phenomenon is not straightforward. The ACS has been used in many studies of this topic. As I write this sentence, the Covid-19 pandemic has cast new doubt on the revival of America's urban areas; many renters are currently fleeing urban areas for cheaper, better places to shelter-in-place. It remains to be seen whether they will return after this pandemic ends.

In an article titled, "Urban revival in America," Victor Couture and Jessie Handbury (2020, p. 1) document, "...after decades of suburbanization, the college-educated population started urbanizing in most large U.S. cities between 2000 and 2010." Their results thus suggest the resurgence of cities (or more accurately, the smaller, downtown areas of cities) actually began prior to 2010. Their focus on downtown areas is important, because some central cities are quite large and feel "suburban" in areas, but the downtown areas that are the focus of the Couture and Handbury (2020) study are entirely "urban" (see also Holian 2019).

Much of the research on America's urban revival has focused on two groups: the college educated and the retired. Kyle E. Walker uses the ACS microdata in his study, "Baby boomer migration and demographic change in US metropolitan areas." Although the popular press has focused on the role of both the young and old in the revival of cities, Couture and Handbury (2020, p. 4) find, "Contrary to claims by the popular press that retiring baby boomers are urbanizing, the older...college-educated groups are still rapidly suburbanizing."

There is substantial nuance in measuring America's urban renaissance. There is also much to celebrate about a renewed interest in urban living. Cities are important for innovation and sustainability. Innovators in close proximity can cross-pollinate each other, while commuters who live near

[13] The statistics in this and the next paragraph come from the following references: On the year 2000 suburbanization rates, Census Bureau (2002, p. 33); on the 2000–2010 annual central city versus suburban growth rates, Frey (2014). The source for the statistics in the next paragraph on the 2010–2019 annual growth rates is Frey (2020).

each other make transit and other green travel options possible. There are also downsides to it. Many groups are locked out of the urban renaissance due to the high real estate prices.

Gentrification is the focus of a study by Lisa Sturtevant (2014), titled, "The New District of Columbia: What Population Growth and Demographic Change Mean for the City." I've read this article many times, and recommend it to anyone who wishes to know more about how to measure gentrification or use the ACS to study internal migration. Washington, DC is also one of my favorite cities in the USA. Growing up, my family avoided cities on our many trips around the country, but we did visit DC a couple of times when my father had conferences there. I'll never forget my first urban experiences that DC provided: riding the metro, eating sushi, and watching the street scene unfold. Sturtevant is not an economist, which for me makes reading her writing refreshing. She is careful to present her analysis to the reader as descriptive and not causal, and she explains details of the data source and, considering her article is a case study, plenty of institutional details.

The beauty of the ACS for studying internal migration is that, for respondents who have moved in the last year, we know where they last lived. Thus we can see where people coming into the city came from, and where people who left the city went to. Sturtevant (2014) finds among whites, 53.7 and 40.1 were in-migrants and out-migrants, respectively. Among blacks, 26.1 were in-migrants but 41.4 were out-migrants. The data she used were from 2006 to 2010, and this period clearly saw many examples of displacement and gentrification.

There's only one sentence of the Sturtevant (2014) study that causes me to bristle every time I read it. As a policy recommendation, the author writes, "...the city can slow the pace of construction of multifamily residential buildings and look for opportunities to encourage a mix of housing types, including townhouses and single-family detached houses." My problem with this recommendation is that, although it might slow gentrification, it seems unlikely to be good for society broadly conceived. In the concluding chapter, I present the framework of Cost–Benefit Analysis (CBA) that can be used to analyze the types of policies that are the focus here; city-level planning departments that approve or deny building permits. There isn't a lot of room for suburban style single-family detached houses in Washington, DC. And where the space for these types of homes exist, it is usually not in transit-friendly and walkable neighborhoods. So, in addition to not solving the underlying problem of gentrification—a

housing shortage—it may also exacerbate climate change. One goal I have for this book is to expose more researchers to the CBA technique, so that when researchers do draw policy recommendations from their work, they are better grounded in a comprehensive, rational decision-making framework.

Another study that makes use of the migration variables in the ACS by Sastry and Gregory (2014) highlights the plight of black Americans in the contemporary period. The authors used responses to the migration questions on the ACS in a study titled, "The location of displaced New Orleans residents in the year after Hurricane Katrina." Sastry and Gregory had access to the restricted version of the data that contains much more precise geographic identifiers than the PUMA-level which is the smallest level available in the public-use version of the data.

The last study I discuss before turning to the next chapter on jobs and school, by John Winters (2017), asks, "Do Earnings by College Major Affect Graduate Migration?" The topic of this study closely connects with not only the migration theme of this chapter, but also the jobs and school theme of the next. Readers of the book all had or will have to decide where to live after their formal education ends. How is what you do and where you live a result of what you studied?

Having access to higher income jobs in a graduate's home state modestly decreases the likelihood of a graduate's leaving. As we have seen above, and like many empirical studies, this one concludes with some policy advice. Like the advice we considered from gentrification, there is no formal accounting for a comprehensive set of impacts. Still, judgment based on likely costs can give some evidence-based suggestions which could be factored in to aid decision-making. Winters cautions, "...providing income subsidies to college graduates to stay and work in state would have a relatively small impact relative to the costs required" (p. 641). Winters concludes with a safe recommendation, "...facilitating the dissemination of better information about local earnings in various majors...may help young people...increase their earnings and their propensity to stay after college." After all, having better income prospects in their home states will help, if not being a magic bullet. It could be virtually costless for states to provide this information.

The next chapter picks up where this chapter ends, with a discussion of how well our education prepares us for the labor market.

Key Terms

Repeated cross-section	Difference-in-differences (D-in-D)	Two-way fixed-effect (TWFE)
Interaction model	Basic D-in-D with control variables	Polynomial
Falsification test		

Questions for Review

1. In Table 3.1, describe the numbers 62.0, 40.6, 36.1, and 31.6. Explain why 16.9 percentage points is a better estimate than 21.4 of the causal effect of TPS on employment among less-educated female Salvadoran migrants. Explain how to use the estimated regression model from the chapter to calculate these statistics.
2. This question revisits the estimates from Table 1.5 from Question 8 at the end of Chapter 1, which examined a subset of lawyers who were business economics and marketing majors. Use the estimates in Column 4 of Table 1.5 to calculate average income for the following four groups of lawyers: (i) male economics majors, (ii) female economics majors, (iii) male marketing majors, and (iv) female marketing majors. Is majoring in economics associated with a higher earnings boost for male or female lawyers? Does the D-in-D have a causal interpretation here? (Hint: Is there a natural experiment?)
3. The case study in Chapter 2 examined building energy codes using regression control. This question asks how we can study the same topic with a D-in-D approach. Modify the file `Script6.R` and calculate average household electricity expenditure, in both California and Texas, for homes built in both the 1960s and 1980s (so, calculate four group averages). Test the hypothesis that the strict energy codes California adopted in 1978 caused electricity consumption, and thus electricity expenditures, to fall.
4. Read the paper by Kuka et al. (2019). Explain how to create a simple difference-in-difference estimate of the effect of the Deferred Action for Child Arrivals (DACA) policy that was discussed in this chapter. The outcome of interest here is teen pregnancy. Describe the estima-

tion subsample (hint: does it contain men of all ages?), the treatment and control group, and the pre-policy and post-policy periods. Use the basic D-in-D regression equation shown below:
$Y_{it} = \beta_0 + \beta_1 TREAT_i + \beta_2 POST_t + \beta_3 (TREAT_i \times POST_t) + u_i$.
Explain how you would construct each of the three variables: Y_i, $TREAT_i$ and $POST_t$.

5. The result here for the effect of Temporary Protected Status (TPS) on Salvadoran immigrants was a basic D-in-D model, while in the original study it was a basic D-in-D model with control variables. Consult the original study and report the estimated effect of TPS on employment among less-educated Salvadoran woman. How does it compare with the result in Table 3.1 of this chapter? Next, report the findings in this study for another outcome (e.g. hourly wage), and finally for another subsample (e.g. more-educated women). Modify the file `script8.R`, replicate the published result, and then reanalyze the data by estimating it as a basic D-in-D model. Describe the natural experiment that separates the individuals into control and treatment, and pre- and post-groups. Does the natural experiment provide a plausible way of assigning treatment as if randomly assigned by experimenters?
6. Another country with a TPS designation is Hondurans. On the USCIS webpage we see that TPS for Honduras required "Continuous Residence in U.S. Since Dec. 30, 1998" and "Continuous Physical Presence in U.S. Since: Jan. 5, 1999." It is thus pretty safe to assume that if a noncitizen immigrant from Honduras was in the country before 1999 they would likely be protected, but if they came in 1999 or after they wouldn't be. Describe how to adapt the file `script8.R` to study the impact of TPS for Honduran migrants, rather than Salvadoran migrants as in the original OZ study. Describe two possible control groups.
7. Two-way Fixed Effect (TWFE) estimators, discussed in footnote 6, are used to analyze policies that start at different times in different locations, while basic D-in-D models analyze policies that start at the same time for all treated entities. Review the data in Kostandini et al. (2013) Table A1 and report the date the first state entity joined the 287(g) program, which permits local and state police to enforce federal immigration law. Describe how you could use the data from the Kostandini et al. (2013) study to do original research using a basic D-in-D framework rather than a TWFE model. This would be an exam-

ple of an event study. Hint: The state of California never joined the 287(g) program, but between 2002 and 2011, four California counties did and they were all in southern California: Los Angeles County Sheriff's Office was first, on 1-Feb-05. San Bernardino County followed on 19-Nov-05, Riverside County on 28-Apr-06, and Orange County on 2-Nov-06. Discuss how a basic D-in-D estimate could be produced with the ACS data from years 2004 and 2007, using data from California counties. Use child-poverty rates as the outcome of the policy (the dependent variable), as in Amuedo-Dorantes et al. (2018).

8. Only one study reviewed in this chapter provides an example of regression control techniques. (The other studies reviewed either used the D-in-D technique, or used regression to estimate descriptive statistics.) Read the Winters (2017) study and compare and contrast the model there with the model from Costa and Kahn (2011) in terms of: fixed effects, whether the dependent variable is binary or continuous, and other control variables used. Describe one precise source of OVB that the model was designed to solve (Hint: see Winters 2017, p. 639).

References

Allen, Treb, Cauê de Castro Dobbin, and Melanie Morten. "Border walls." No. w25267. National Bureau of Economic Research, February 2019.

Amuedo-Dorantes, Catalina, Esther Arenas-Arroyo, and Almudena Sevilla. "Immigration enforcement and economic resources of children with likely unauthorized parents." Journal of Public Economics 158 (2018): 63–78.

Callaway, Brantly, and Pedro H.C. Sant'Anna. "Difference-in-differences with multiple time periods." Journal of Econometrics (in press, 2020). https://doi.org/10.1016/j.jeconom.2020.12.001.

Census Bureau. "Demographic trends in the 20th century." Census 2000 Special Reports, 2002. https://www.census.gov/prod/2002pubs/censr-4.pdf.

Costa, Dora L., and Matthew E. Kahn. "Electricity consumption and durable housing: Understanding cohort effects." American Economic Review: Papers & Proceedings 101, no. 3 (2011): 88–92.

Couture, Victor, and Jessie Handbury. "Urban revival in America." Journal of Urban Economics 119 (September 2020): 103267.

Frey, William H. "Will this be the decade of big city growth?" May 23, 2014. https://www.brookings.edu/opinions/will-this-be-the-decade-of-big-city-growth/.

Frey, William H. "American cities saw uneven growth last decade, new census data show." May 26, 2020. https://www.brookings.edu/research/new-census-data-show-an-uneven-decade-of-growth-for-us-cities/.

Goodman-Bacon, Andrew. "Difference-in-differences with variation in treatment timing." Working Paper. July, 2019.

Holian, Matthew. "Where is the city's center? Five measures of central location." Cityscape: A Journal of Policy Development and Research 21, no. 2 (2019).

Kostandini, Genti, Elton Mykerezi, and Cesar Escalante. "The impact of immigration enforcement on the US farming sector." American Journal of Agricultural Economics 96, no. 1 (2013): 172–192.

Kuka, Elira, Na'ama Shenhav, and Kevin Shih. "A reason to wait: The effect of legal status on teen pregnancy." AEA Papers and Proceedings 109 (2019): 213–217.

Orrenius, Pia M., and Madeline Zavodny. Beside the golden door: US immigration reform in a new era of globalization. AEI Press, 2010.

Orrenius, Pia M., and Madeline Zavodny. "The impact of temporary protected status on immigrants' labor market outcomes." American Economic Review: Papers & Proceedings 105, no. 5 (2015): 576–580.

Sastry, N., and J. Gregory. "The location of displaced New Orleans residents in the year after Hurricane Katrina." Demography 51, no. 3 (2014): 753–775.

Sturtevant, Lisa. "The new District of Columbia: What population growth and demographic change mean for the city." Journal of Urban Affairs 36, no. 2 (2014): 276–299.

Walker, Kyle E. "Baby boomer migration and demographic change in US metropolitan areas." Migration Studies 4, no. 3 (2016): 347–372.

Winters, John V. "Do earnings by college major affect graduate migration?" The Annals of Regional Science 59, no. 3 (2017): 629–649.

CHAPTER 4

Paying the Bills: School, Jobs, and Health Insurance

Many Americans spend a significant portion of their lives in school, or worrying about their children's education. Schooling is so important because there are a lot of bills to pay. Health care costs loom especially large, and the Affordable Care Act (ACA, also known as Obamacare) is a controversial contemporary policy. Is it merely providing a more robust social safety net, or is it a slippery slope toward socialism? This chapter describes a replication of a study that examines the effect of the ACA on job lock and entrepreneurship.

This chapter also describes some of the extensive research that labor economists have carried out using the ACS. We'll see the D-in-D technique is used in many of these studies. But first, we revisit the following question, what is the right major to study in college?

Chapter 1 discussed a study that used the ACS data to calculate the average earnings of lawyers by college major. Chapter 2 discussed how, once a study is replicated, it is often straightforward to extend it in a way that creates new knowledge. In this chapter, we revisit the lawyer earnings by college major study, and use a modified version of the replication file to determine the most popular and highest paid majors in a different occupation, software developers. The file (`script9.R`) generalizes the lawyer earnings analysis, and enables a user to automate calculation of most popular and highest paid majors for any industry.

M. J. Holian, *Data and the American Dream*,
https://doi.org/10.1007/978-3-030-64262-4_4

Table 4.1 Most popular majors and average earnings for software developers

Major	*% of developers*	*Mean earnings*
Computer Science	29.1	94,075
Electrical Engineering	10.6	101,632
Computer Engineering	8.2	95,116
Computer and Information Systems	4.8	77,352
Mathematics	3.9	101,109
Business Management and Administration	3.0	81,473
Mechanical Engineering	2.8	97,085
General Engineering	2.5	88,590
Physics	2.2	104,122
Management Information Systems and Statistics	2.1	87,078
General Business	1.9	88,970
Information Sciences	1.5	86,075
Electrical Engineering Technology	1.2	82,637
Economics	1.2	93,357
Accounting	1.1	89,068
Psychology	1.0	82,243
Biology	1.0	85,886
English Language and Literature	0.9	77,270
Civil Engineering	0.8	103,391
Finance	0.8	91,382

I focus on software developers because among the skills this book illustrates is computer programming. Programming skills, with a knowledge of causal inference techniques and human behavior, is a powerful combination for success in currently hot data analytics careers. It's a path too few of my students take, and one of my goals in writing this book is to help more students use their economics education to find jobs in it and in the lucrative software development industry that flourishes in Silicon Valley and many other areas of the country.

Table 4.1 shows the top 20 college majors among software developers. Nearly a third (29.1%) of software developers were computer science majors. This is not a surprise. However, at $94,075 computer science majors do not have the highest average earnings. (In fact, physics majors have the highest average earnings among the top 20 at $104,122.)

Table 4.1 also reveals that English language and literature is a top 20 major among software developers, which might come as a surprise to some readers. Another possibly surprising finding is that social science majors—economics and psychology—are represented in the top 20 majors, at 1.2%

and 1% of developers, respectively. Finally, I note that economics majors do quite well as developers. With average earnings of $93,357, economics majors earn as much on average as computer science majors.

Although average income differences across majors reflect both selection and treatment effects, the strong earnings of economics majors as programmers is consistent with a world where having a social science perspective is good preparation for jobs in the technology sector. It is true that the typical computer science major comes out of school with better programming knowledge than the typical economics major. However, economics and other social sciences are more likely to emphasize causal inference, decision-making, and human behavior, while exposing students to the way programming is used in economic research.

I turn now to a review of labor economics research that used the ACS microdata.

The killing of George Floyd at the hands of a Minneapolis police officer in May of 2020 sparked nationwide protests and reflection on the unequal treatment of blacks in American society. The phrase "Black Lives Matter" emerged as a contemporary social controversy. American economists, who are part of the society they study, have also embarked on a deeper reflection of whether they have done enough through their research to advance social justice.[1]

One study that uses the ACS to study labor-market discrimination is titled, "Changes in the earnings of Arab men in the US between 2000 and 2002," by Alberto Davila and Marie Mora (2005), hereafter DM. Using the D-in-D techniques that we saw in the last chapter, DM studied the impact of 9/11 terrorist attacks on labor-market outcomes of Arab immigrants. They find that hourly wages of Middle Eastern Arab Men fell from $16.37 in 2000, before 9/11, to $13.77 in 2002. Is this evidence of discrimination? Maybe, but the authors point out that the fall in average earnings could have been due to the macroeconomic slowdown that also took place during this period. In other words, a difference in means by itself is not a compelling estimate of the causal effect of discrimination on earnings.

[1] See, Amara Omeokwe, "Economics Journals Faulted for Neglecting Studies on Race and Discrimination," Wall Street Journal, July 12, 2020.

To account for this, DM use a control group of US-born non-Hispanic white men. They find for this group that hourly wages actually rose from $16.47 in 2000 to $17.03 in 2002, despite the economic slowdown.[2] With these four averages, one can calculate the *basic difference-in-differences* (D-in-D) estimate defined in the previous chapter, and they also find results that are similar in magnitude in a basic D-in-D model with control variables, leading the authors to conclude (p. 587), "Our interpretation is that the unanticipated events of September 11th, 2001 negatively affected the labor-market income of the groups most closely associated with the ethnicity of the terrorists." This is a clear causal claim. With these results we cannot say whether and to what extent discrimination was to blame, or the precise mechanisms through which discrimination operates, but the results are certainly consistent with a world where workers from Arab countries saw their job prospects diminished after 9/11.

The next three studies deal more directly with public policy. Two studies by David Sjoquist and John Winters examine a state-level public policy, referred to as "merit aid," which is designed to increase the stock of educated workers in a state, by offering scholarships that are tied to good performance, for example, with minimum grade point average (GPA) requirements. The concept of "Brain Drain" refers to a phenomenon where highly educated professionals leave their home states or countries, lowering productivity in the areas they leave. I was born in Ohio, and all of my degrees are from public universities in the state. After finishing my Ph.D. at Ohio State University, I migrated to California where I now use my training to teach college students there. Does it make sense for Ohio taxpayers to fund higher education for people like me, who leave the state after graduating?

The study by David Sjoquist and John Winters titled, "State merit-based financial aid programs and college attainment" examines college completion. Their results do not indicate there is a statistically significant positive effect of these policies on college completion rates. Their other study is titled, "State merit aid programs and college major: A focus on STEM." The focus in both is on college major. They find evidence that merit-aid

[2] As we saw in Chapter 2, it is common for researchers to work with logged values of variables, rather than levels. Table 1 of their study reports the average of the natural log of hourly wages, which I converted to levels. The figures they reported were, for Middle Eastern Arab Men, 2.796 and 2.623, and for US-born non-Hispanic white men: 2.802 and 2.835. I converted these to levels with the equation $e^{2.796} = 16.37$, where e is approximately equal to 2.718.

programs make students less likely to pursue STEM majors, a result that could be explained by a situation where students avoid difficult majors to avoid losing their GPA-dependent merit aid.

A third labor-related study by Robert Thornton and Edward Timmons titled, "Licensing One of the World's Oldest Professions: Massage" used the ACS microdata to study the impact of occupational licensing laws on wages of massage therapists. Most states require workers in at least some occupations to have a license. Hair dressers are a good example, almost all states require hair dressers to have a license from a state board of cosmetology. One state licenses florists. Do these laws protect consumers from unqualified service providers, or do licensing laws mainly serve to prevent competition among producers and raise prices for consumers? Among the findings in Thornton and Timmons (2013) is that the adoption of licensing caused earnings of massage therapists to rise.

The studies by Thornton and Timmons (2013), and Sjoquist and Winters (2015a, b) all utilize the two-way fixed effect (TWFE) model, introduced in the previous chapter. These studies also make use *merged data*, in both cases state-level: licensing laws for massage therapists, or the existence of merit-based scholarships programs. All are certainly candidates for beginning students to replicate, but as discussed previously TWFE models are difficult to interpret correctly.[3] Two end-of-chapter review questions present some ideas for using the ACS data, and the merged data shared by these authors, to estimate event study models, using the powerful basic D-in-D model.

The next section describes a study of the ACA using a basic D-in-D model, and a twist on it that expands on the idea of D-in-D with control variables that we saw in the last chapter.

Health Insurance

The Affordable Care Act (ACA) of 2010, widely known as ObamaCare, had several provisions, including: a dependent coverage provision, expansion of state Medicaid programs, an individual mandate, an employer mandate,

[3] All three studies include the data on state-level policies in tables, which is fortunate because authors that use merged data often don't include it in tables in their published articles, or archive their research data.

a Health Insurance Marketplace, and a community rating, to name six.[4] Several studies have used the ACS data to study various provisions of the ACA.[5]

One goal of the ACA was to facilitate entrepreneurship by reducing a worker's dependency on employer-provided health insurance. The study by Bailey and Dave (2019), "The Effect of the Affordable Care Act on Entrepreneurship among Older Adults" is the focus of this section. I'll describe their approach using the basic D-in-D model introduced in the last chapter, and then discuss the generalized version of it that they actually estimated.

Bailey and Dave (2019, p. 143) write, "...many of the main ACA provisions took effect in January 2014..." and they therefore select 2014 as the year it went into effect. They note that the various provisions work both to encourage and discourage self-employment. For example, the employer mandate may make self-employment less attractive, while the Health Insurance Marketplace may make self-employment more likely.[6]

Bailey and Dave (2019, p. 141) focus on older adults, "...whose higher average health costs and health insurance premiums make health insurance more salient to their labor market decisions." They use workers aged 65–69 who are already eligible for Medicare as a control group that was less affected by the ACA. As a treatment group they use workers aged 60–64 that are similar in age but should have been more affected by the ACA as they are not eligible for Medicare. The main finding in the study is that by, "...lowering the cost of non-employer health insurance policies to older adults, the ACA appears to have eased their transition from employment to self-employment."

[4] The dependent coverage provision mandated that children can remain on the health insurance plans of their parents until age 26. The expansion of state Medicaid programs affected low income households and provided partial federal funding to states. The individual mandate required most individuals to have health insurance, or face a tax penalty, although under Trump, the penalty was eliminated beginning in 2019. The employer mandate required all employers with at least 50 employees to offer health insurance to full-time employees. The Health Insurance Marketplace established health insurance exchanges and subsidies for low and moderate income households.

[5] Condliffe et al. (2017), Frean et al. (2017), Lee and Winters (2020), Dillender (2014). Some of these studies use TWFE models.

[6] As a result, the effect on entrepreneurship they estimate can be thought of as the net effect of the ACA. In their article they refer to the net effect as a "reduced form effect" and contrast this with a "structural effect," which would measure some specific policy mechanism.

Before presenting results, it is common for D-in-D studies to support their selection of an event as a natural experiment by presenting an analysis of *pre-trends*. Here this involves demonstrating that the self-employment rates for the control and treatment groups moved in parallel before treatment. This presentation is sometimes called a test for parallel trends, but in fact it is usually just a visual inspection of a chart like Fig. 4.1.

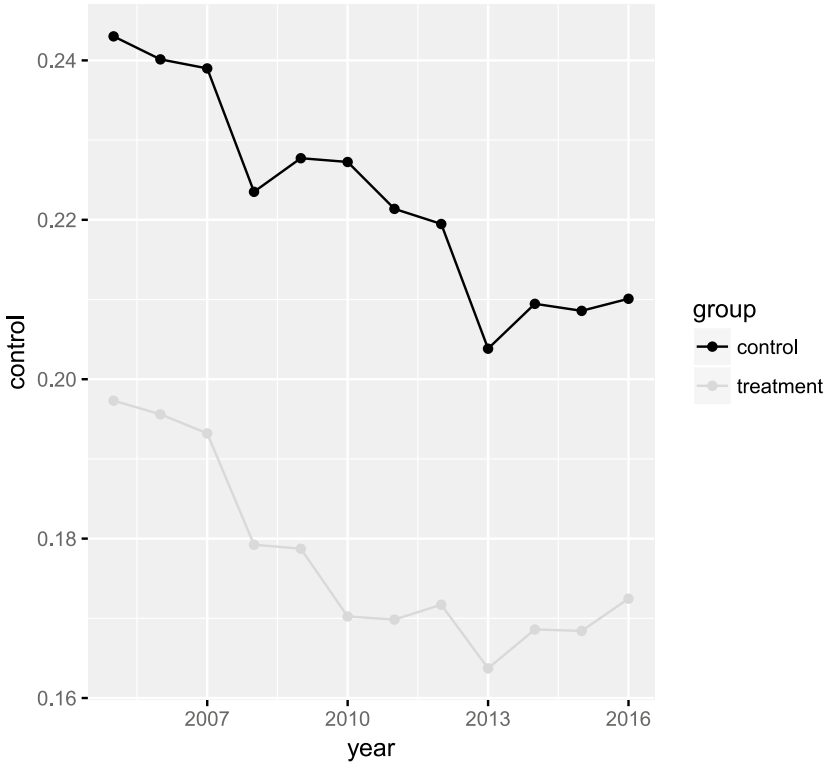


Fig. 4.1 Self-employment trends, treatment and control groups, ACS 2005–2016

Figure 4.1 shows that, although self-employment rates are higher for the older cohort, they seemed to have largely though not perfectly moved in parallel over the 12 years shown in the figure. The key part of the analysis is what happened to self-employment after 2014. We see in the figure it

Table 4.2 Self-employment rates among two groups of older works, pre- and post-ACA

	Age 65–69	*Age 60–64*
Pre-ACA (surveyed 2005–2013)	18	14.7
Post-ACA (surveyed 2014–2016)	16.9	14.1
Difference	1.1	0.6
Difference in differences	1.1 − 0.6 = 0.5	

went up fairly sharply for the treatment group, which is consistent with a world where ACA encouraged self-employment among those too young to qualify for Medicare. An increase in self-employment is also evident in the control group after 2014, but it is not as steep.

Next, consider the numbers in Table 4.2. This table shows self-employment rates among individuals in two age cohorts (a younger group not eligible for Medicare, aged 60–64 and an older group that has always been eligible for Medicare, aged 65–69) before and after 2014 (the year most ACA provisions went into effect).

The self-employment rate among workers aged 60–64 before most provisions of the ACA went into effect was 14.7%, and this proportion fell to 14.1% after ACA, a reduction of 0.6, or just over half a percentage point. From this difference in means alone, it would seem the ACA did not succeed in raising entrepreneurship rates. However, consider what happened to the older group of workers in this sample. Their self-employment rate fell from 18 to 16.9%, a reduction of 1.1 percentage points. It is possible the ACA had nothing to do with this reduction, however, because being over 65 these workers were eligible for Medicare throughout the entire time period under consideration. Perhaps macroeconomic factors like interest rates and technology made self-employment less likely for everyone. In light of this, it is possible the ACA mitigated the reduction in self-employment; in other words, self-employment rates would have been even lower among the group aged 60–64 who were surveyed between 2014 and 2016 if not for the ACA.

If the ACA did not affect the insurance market for this younger age group, we might have expected self-employment rates to mirror those of the older age group, namely, the self-employment rate would have been 1.1 percentage points lower among the younger group. In this case, the self-employment rate among workers aged 60–64 would have been 14.7

minus 1.1, or 13.6%. Instead, it was 14.1%, or half a percentage point higher. The difference-in-difference estimate shown in the bottom line of Table 3.1 is thus 0.5. It looks like the ACA raised self-employment by half-a-percentage point among workers aged 60–64.

We see in the bottom row of Table 4.2 that the D-in-D estimate is just the difference between two differences in means. All we have to do to calculate it is take the difference in means for two groups, and subtract them. Again, no fancy concepts other than averages and subtraction are needed to calculate it.

We also saw in the last two chapters that multivariate regression offers a convenient way of calculating group averages. The file `script10.R` estimates the model shown below, using a sample of older workers between the ages of 60 and 69, and data from the 2005–2016 ACS samples:

$$SELFEMP_{it} = \alpha + \beta AGE6064_i + \gamma POST_t + \delta(AGE6064_i \times POST_t) + \varepsilon_{it},$$

where the dependent variable $SELFEMP_i$ is a binary variable equal to zero if individual i works for an employer, and equal to one if they are self-employed. The variables on the right-hand side are also binary variables. $AGE6064_i$ is equal to one if individual i is between the ages of 60 and 64 and zero if they are in the older age group, and $POST_t$ is equal to one if the individual was sampled in 2014, 2015, or 2016, which is the period after most of the ACA went into effect. $POST_t$ is equal to zero for individuals sampled before 2014. The third variable on the right-hand side is the product of $AGE6064_i$ and $POST_t$. This is the interaction term, as we saw in the last chapter, and the product of these two binary variables will be equal to one only if both are one, and it is zero otherwise.

When we estimate this equation, we find values for the coefficients (α, β, etc.) The estimated equation is shown below:

$$SELFEMP_{it} = 0.18 - 0.033 AGE6064_i - 0.011 POST_t + 0.005(AGE6064_i \times POST_t).$$

As we saw before, we can interpret this equation by plugging in actual values for the two independent variables. Because both $AGE6064_i$ and $POST_t$ can take on only two values (zero or one) so there are four combinations, which produce the four averages we see in Table 4.2. For example, plugging in values of zero for the variables on the right-hand side, the predicted value is 0.18. Thus we predict someone in the 65–69 cohort has an

18% chance of being self-employed in the pre-ACA period. The reason we find this is because 18% of people aged 65–69 in the 2005–2013 sample were self-employed.

Like Orrenius and Zavodny (2015), BD estimate the D-in-D model using a set of control variables, but here there is another twist. In the BD model, they include dummies for each year, rather than a single $POST_t$ variable as in my basic D-in-D adaptation of their study. One reason they do this is because they were worried that the effect of the Great Recession could bias their results. The 2005–2013 period, which is the pre-ACA period in the BD sample, spans both before and after the recession that began in 2008. They also include binary variables for each age, rather than a binary for the group. However, the interaction term is the same term as in the basic D-in-D model.[7]

Formally, *the fixed effect D-in-D* model (with control variables) is written as follows:

$$\begin{aligned} SELFEMP_{ist} = {} & \alpha_{FE} + \delta_{FE}(AGE6064_i \times POST_t) + \\ & + \sum_{a=61}^{69} \beta_a AGE_{ai} + \sum_{k=Alaska}^{Wyoming} \beta_k STATE_{ks} \\ & + \sum_{j=2006}^{2016} \gamma_j YEAR_{jt} + \beta_1 MALE_i + \beta_2 ASIAN_i \\ & + \beta_3 BLACK_i + \beta_4 HISPANIC_i + \beta_5 MARRIED_i \\ & + \beta_6 NCHILD_i + \beta_7 HSGRAD_i + \beta_8 COLGRAD_i \\ & + \beta_9 STEMP_{st} + \varepsilon_{ist}. \end{aligned}$$

Although this equation looks more intimidating than the basic D-in-D model shown above, other than the fact that there are more independent variables, nothing is fundamentally different about interpreting the main variable of interest. They end up finding the D-in-D estimate, the estimate of the δ_{FE} coefficient, at 0.572, is nearly the same in this fixed-effect D-in-D model as in the basic D-in-D model I presented above (where it was 0.5). We wouldn't have known this was the case without going through the

[7] Michael Bailey's textbook *Real Econometrics*, section 8.5, labels this type of model *the fixed effects D-in-D* estimator. It bears resemblance to the TWFE estimator but, like basic D-in-D, it includes an interaction term. In TWFE studies like Condliffe et al. (2017) the main variable of interest switches on and off at different locations at different times; it is not an interaction term. The policy variable in Bailey and Dave (2019) starts at the same time for all individuals.

trouble to estimate the elaborate model, but the fact that estimates from both models are similar gives us more confidence the findings are not just a result of modeling choices.

Key Terms

Basic difference-in-differences (D-in-D) Merged data Pre-trends analysis Fixed-effect D-in-D

Questions for Review

1. Compare the estimate from the basic D-in-D model presented in this chapter with the fixed effect D-in-D model estimated by Bailey and Dave. How much do the results differ?
2. Explain how to create a simple D-in-D estimate of the effect of 9/11 on labor-market discrimination of Arab immigrants. The outcome of interest here is employment. Earnings are another outcome of interest. Describe the estimation subsample. Use the basic D-in-D regression equation shown below:
   $Y_{it} = \beta_0 + \beta_1 TREAT_i + \beta_2 POST_t + \beta_3 (TREAT_i \times POST_t) + u_i$,
   and explain how you would construct each of the three variables: Y_i, $TREAT_i$ and $POST_t$.
3. Two-way Fixed Effect (TWFE) estimators are used to analyze policies that start at different times in different locations, while D-in-D models analyze policies that start at the same time for all treated entities. Review the data in Sjoquist and Winters (2015a) Table 1 and report which state was the first to adopt a merit aid scholarship program. Describe how you could adapt the Sjoquist and Winters (2015a) study using a basic D-in-D framework rather than a TWFE model. Hint: The state of Kentucky adopted a merit aid program in 1999, but Tennessee didn't adopt one until 2003. Discuss how a basic D-in-D estimate could be produced with the ACS data from years 2009–2011. Use college completion and STEM major as the outcomes of interest, as in Sjoquist and Winters (2015a, b).

4. Another example of a study that uses a TWFE estimator that can be turned into basic D-in-D models is the study by Thornton and Timmons (2013). Ohio has licensed massage therapists since 1915 while Michigan didn't adopt the requirement until 2011. Describe how we would set this up as a basic D-in-D model.
5. A study by James Bailey (2017) also used a D-in-D approach, this time focusing on younger workers and the impact of the dependent coverage provision. Describe the control and treatment groups, and the year treatment went into effect for the purposes of the study.

References

Bailey, James. "Health insurance and the supply of entrepreneurs: New evidence from the affordable care act." Small Business Economics 49, no. 3 (2017): 627–646.

Bailey, James, and Dhaval Dave. "The effect of the Affordable Care Act on entrepreneurship among older adults." Eastern Economic Journal 45, no. 1 (2019): 141–159.

Condliffe, Simon, Matt B. Saboe, and Sabrina Terrizzi. "Did the ACA reduce job-lock and spur entrepreneurship?" Journal of Entrepreneurship and Public Policy 6, no. 2 (2017): 150–163.

Davila, Alberto, and Marie T. Mora. "Changes in the earnings of Arab men in the US between 2000 and 2002." Journal of Population Economics 18, no. 4 (2005): 587–601.

Dillender, Marcus. "Do more health insurance options lead to higher wages? Evidence from states extending dependent coverage." Journal of Health Economics 36 (July, 2014): 84–97.

Frean, Molly, Jonathan Gruber, and Benjamin D. Sommers. "Premium subsidies, the mandate, and medicaid expansion: Coverage effects of the Affordable Care Act." Journal of Health Economics 53 (2017): 72–86.

Lee, Jun Yeong, and John V. Winters. "State medicaid expansion and the self-employed." IZA Discussion Paper No. 12997 (2020).

Orrenius, Pia M., and Madeline Zavodny. "The impact of temporary protected status on immigrants' labor market outcomes." American Economic Review: Papers & Proceedings 105, no. 5 (2015): 576–580.

Sjoquist, D. L., and J. V. Winters. (2015a). "State merit aid programs and college major: A focus on STEM." Journal of Labor Economics 33, no. 4: 973–1006.

Sjoquist, D. L., and J. V. Winters. (2015b). "State merit-based financial aid programs and college attainment." Journal of Regional Science 55, no. 3: 364–390.

Thornton, Robert J., and Edward J. Timmons. "Licensing one of the world's oldest professions: Massage." The Journal of Law and Economics 56, no. 2 (2013): 371–388.

CHAPTER 5

Home Economics: Family Matters

This chapter describes studies of marriage and family. There is understandably a great deal of interest in these topics among many of my 20-year-old students. Family issues are deeply connected with economic questions addressed in earlier chapters, including migration, labor supply, and entrepreneurship. We'll see how the ACS can be used to describe the rise of the gig economy, the difficulty of studying the causal effect of children on a family's outcomes, and a possible way to measure the causal effect of the business cycle on fertility.

Let's begin by revisiting an earlier example of family economics. In Chapter 1 we saw a regression model that showed families where the children have the same gender live, on average, in homes with slightly fewer bedrooms. Is this a treatment or selection effect? I argued there, it should be interpreted as a treatment (or causal) effect, precisely because families don't select the gender of their children and it is instead determined by nature. But in fact, it's not always true that families don't select the gender of their children. Families that adopt may be given a choice. And families that use sex-selection practices (selective abortion) can also control the gender of their children.

I stand by my interpretation of the finding of slightly smaller average number of bedrooms among families with same gender children as a causal effect, because research using the ACS by Blau et al. (2020) indicates there is no longer evidence of son preference in the USA. In many developing

M. J. Holian, *Data and the American Dream*,
https://doi.org/10.1007/978-3-030-64262-4_5

countries, son preference is widely reported. In places like these, where more couples (though certainly not all or even most) do actively control the gender of their children, we would have to worry about the selection effect in interpreting a regression model like the one we saw in Chapter 1. While there's no question on the ACS about whether a respondent prefers boy or girl children, Blau et al. (2020) find that couples whose first child was a girl are no more likely to have a second child, and there is no evidence that a woman is more likely to be a single-mom if her first child was a girl. Earlier research by Dahl and Moretti (2008) found the opposite using earlier data, so I interpret Blau et al.'s (2020) results as, there is no longer any evidence of son preference in the USA, at least, in a way that would prevent interpreting the difference in means we found in the bedrooms-sibling gender regression model as a causal effect.[1]

Using the ACS and a D-in-D strategy, Argys and Averett (2019) find higher levels of education for Chinese migrants born after China's one child policy, compared with migrants from other East Asian countries. This finding suggests parents trade-off quality and quantity in fertility decisions.

Another family economics topic is poverty among same-sex couples. Schneebaum and Lee Badgett (2019) finds higher poverty rates among gay and lesbian households. Although the ACS doesn't ask about sexual preferences, it is possible to infer someone is homosexual if the person classified as their partner is of the same gender. Identifying these households would require some intermediate programming abilities (that are demonstrated in `script4.R`) but fortunately, IPUMS makes identifying them easier by creating a variable SSMC (same-sex married couple).[2] The work by Schneebaum and Lee Badgett (2019) connects the marriage and family theme of this chapter to the discussion of discrimination in labor markets in Chapter 4. Although largely descriptive, these results are consistent with a world where gays and lesbians face discrimination, even if their results don't indicate the precise ways in which discrimination causes higher poverty rates.

These are just a few examples of studies that have used the ACS to study questions of marriage and family. Sometimes it may seem these questions

[1] It is possible to view Blau et al. (2020) as a replication of Dahl and Moretti (2008), using the terminology developed in Chapter 2.

[2] From the IPUMS webpage, "SSMC reports whether the head of household and spouse are a same-sex married couple. Beginning in the 2013 ACS/PRCS, same-sex married couples are included in the 'married spouse present' category..." And in a sign of the times, "Prior to the 2013 ACS/PRCS, same-sex married couples were recoded by the Census Bureau from married to unmarried partners."

are outside the domain of economics as a discipline. I strongly disagree with this sentiment. What happens in our personal lives is deeply connected to the economic choices we make, thus family economics cannot be ignored. The next chapter provides an example that highlights this, in estimating the impact of urban form on vehicle ownership.

The remainder of this chapter contains two sections. The first section begins with some results from the ACS regarding growth in employment in ridesharing, gives an example of the intangible benefit for families of the flexibility that comes with self-employment, and describes a study that illustrates how difficult it is to use *regression control* techniques to estimate causal effects in questions of family. The last section presents a case study of research that uses a *difference-in-differences* model to estimate the impact of recessions on fertility.

Marriage, Self-Employment, and the Gig Economy

The rise of the "gig economy" refers to the growth of people working for companies like Uber, Lyft, Instacart, Task Rabbit, etc. What these firms have in common is that they are all platforms for exchange, enabled by technology, rather than firms in the traditional sense (Munger 2018). Workers for these firms are not traditional employees and work as little or as much as they want.

Some recent research has cast doubt on whether surveys like the ACS can be used to study the gig economy. In an article titled, "The Rise of the Gig Economy: Fact or Fiction?" Abraham et al. (2019) write that, "Core household surveys...appear not to be capturing changes that other data sources tell us are occurring" (p. 360). Evidence from various proprietary data sources, like Uber corporate data, data on bank deposits from ridesharing companies, and tax records, paints a picture of a rapid rise, while, "In contrast to these data sources, the CPS does not capture the rapid rise in gig activity in the passenger transportation sector" (p. 359).

The CPS, or Current Population Survey, is one of the large core household surveys, but it is not the largest. Abraham et al. (2019) do not discuss the ACS, which surveys far more people than the CPS. In `script11.R` I use data from individuals working in occupation code 809 (Taxi cab drivers and chauffeurs) and calculate the fraction who are self-employed versus salaried employees, each from 2005 to 2018. In Fig. 5.1 we see a noticeable increase in the fraction of workers who are self-employed, starting

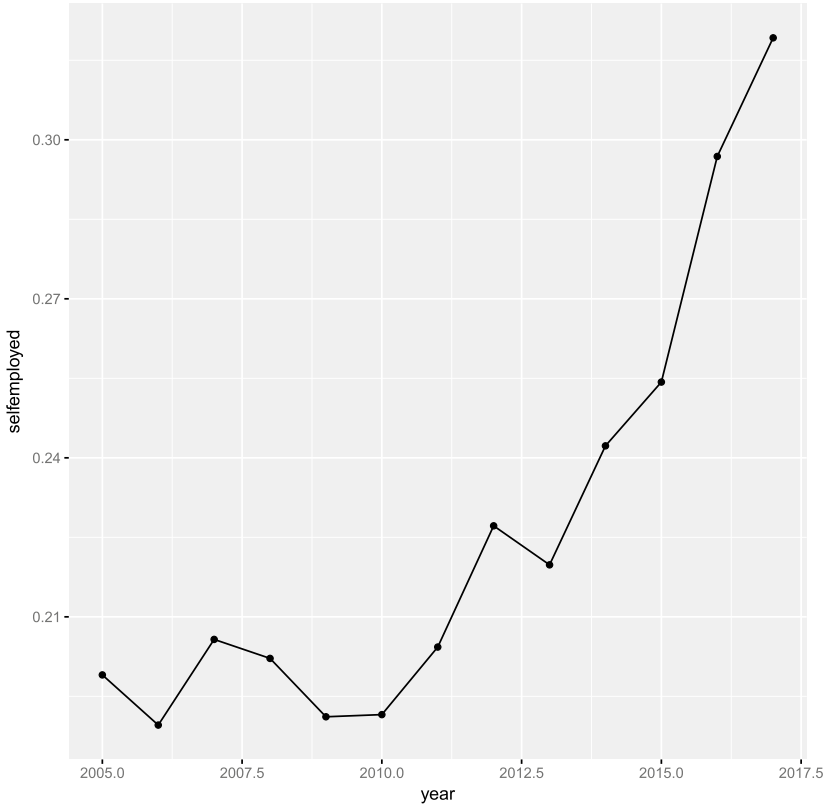


Fig. 5.1 Proportion of taxi drivers and chauffeurs who report being self-employed

in 2012 when Uber began operations in earnest. In a California sample, not shown in the figure, the fraction nearly doubles from 15%, where it had been since 2005, to nearly 30%. In the national sample used in Fig. 5.1 the increase is smaller but still very noticeable.

The ACS is not able to measure many aspects of the rise of the gig-economy. However it is feasible to study the gig economy with the ACS in some settings, and the ACS seems particularly well-suited to study the connections between gig work and family issues, because we know so much about the households surveyed in the ACS.

A healthy economy will not only provide all workers with jobs, but will also provide the type of employment that best suits their lifestyles. For families with a young child, like my own, having flexible hours allows parents to dedicate more time to parenting. When our daughter was born, my wife Bridget was a director at a nonprofit organization. When she returned to work after maternity leave, her employer accommodated her with a somewhat more flexible schedule, but after a few months she decided to resign and start her own nonprofit consulting firm. In quitting her job and striking out on her own, Bridget took a risk, but it has paid off allowing her to earn income while setting her own hours.

There are many factors that led to Bridget's decision to start her own business, including being raised in an entrepreneurial household, her M.B.A., and her own business organization abilities, but clearly childrearing was also a factor. I am a tenured college professor and this also probably helped, as it gives us a steady (if not very high) income, along with health care benefits. The study by Bailey and Dave (2019) discussed in the last chapter examined the effect of health insurance on self-employment, and found evidence of a causal connection between the two.

A study by Maria Marshall and Anna Flaig titled, "Marriage, Children, and Self-Employment Earnings," uses the ACS to study the interplay between these sorts of family issues and entrepreneurial success, by examining the effect of marriage and children on earnings among self-employed women. This study lays some important groundwork for future study of these topics. It also a useful example of the challenges of the *regression control* technique. Marshall and Flaig's (2014, hereafter MF's) baseline model is the one shown below, using one year of ACS data, from 2009:

$$INCEARN_i = \alpha^s + \beta^s MARRIED_i + e_i^s.$$

The key coefficient is β^s. Because the model above is a bivariate regression with a binary independent variable, the estimate of β^s will be equal to the difference in mean income between married and single women. The estimation subsample here consists only of self-employed women, with unincorporated businesses (like sole proprietorships) between the ages of 22 and 65.

Before I reveal what MF found when estimating this equation, take a moment to ask yourself what you'd expect to find. Historically, marriage has often led to a division of labor in the household, with husbands taking on more work responsibilities, and wives more household responsibilities.

This "specialization hypothesis" predicts marriage will cause earnings to rise for men and fall for women. If you reviewed the previous studies of the effect of marriage on male earnings, you would find many studies that have found married men earn more on average than never-married men. This could be because marriage actually causes higher earnings through the division of labor of home and work responsibilities. But it could also be true that men with higher earnings potential are more attractive marriage partners. In other words, the findings in the literature reflect both treatment and selection effects.

Focusing on the effect of marriage for women, MF find married women do not earn more on average; average earnings are $23,754 for single and slightly lower at $23,542 for married women, though this difference is not statistically significant.[3] The difference in means of $212 is a descriptive statistic and doesn't tell us how a single woman should expect to see her income change if she gets married, thus MF employ a regression control technique to better estimate the causal effect of marriage on earnings.

MF estimate multivariate regression models with many control variables. To illustrate their approach, I'll focus on just one control variable, AGE. It would be important to control for age if married women are older on average and thus have more experience and higher earnings as a result. In such a case, AGE meets the two Omitted Variable Bias conditions: Age (1) explains earnings, and (2) is correlated with marriage. Regression control allows us to control for the effect of age simply by including AGE on the right-hand side:

$$INCEARN_i = \alpha^l + \beta^l MARRIED_i + \gamma AGE_i + \delta AGE_i^2 + u_i^l.$$

We see in this equation that both AGE and the squared value of it are included as two control variables. Including a squared value of a variable is like including a log transformation of a variable, as we saw in Chapter 2, in that both are ways of estimating nonlinear effects.[4] Controlling for AGE,

[3] Recall as discussed in Chapter 1, the coefficient on the constant term α^s will be equal to the average income for the zero group (the group for which the binary variable is zero); thus the estimated α^s coefficient is 23,754. Meanwhile, the estimated β^s coefficient is equal to the difference in means, $23{,}542 - 23{,}754 = -212$. I calculate the β^s coefficient using average income statistics reported in Table 1 of Marshall and Flaig (2014).

[4] Models with squared, cubed, or higher order terms are called *polynomial models*. Polynomial models are harder to interpret than log models, but are more flexible in the type of nonlinear relationships they are able to measure. In interpreting polynomial models, you can

we would expect the estimate of β^l to be less biased than that of β^s in the previous equation. However, β^l could still be a very biased estimate of the causal effect of marriage on earnings, as there are many other factors that are not included in my equation above.

MF do include numerous control variables beyond AGE in the models they estimate (found in Table 3 of their article), but there are some variables that they could not measure and thus could not include. One difficult factor to control for is a woman's preference for career versus family life. It is not even clear to me how this can be measured, but, it is probably true that women for whom career is very important will be less likely to be married, and also tend to have fewer children than women for whom career is not as important. Thus the coefficient on marriage will pick up this selection effect, as well as the treatment effect.[5]

Not controlling for enough factors (like preferences for career versus family) illustrates a common limitation of regression control models, as we also saw in Chapter 2 in our discussion of the Costa and Kahn (2011) study. The MF study also illustrates a danger in controlling for *too many* factors. Consider homeownership, which they control for with a binary variable. If being a homeowner means, all else equal, a woman has more financial capital, then this could make self-employment more likely and more lucrative. It is also typically easier for married couples to afford homes (due to dual incomes, sharing a bed, etc.). In other words, marriage makes homeownership more likely, and homeownership makes self-employment more likely and more lucrative. Homeownership is a channel through which marriage potentially influences both self-employment and earnings.

In this example, homeownership illustrates what Angrist and Pischke (2014, pp. 214–217) call a *bad control*, and what Bailey (2017, p. 157) calls a *posttreatment variable*. Although homeownership satisfies the two OVB conditions, it is itself an outcome related to the variable of interest and thus should not be used as a control. It is difficult to intuitively characterize all the requirements for control variables in a regression control strategy.

always fall back on finding two fitted values by plugging in two sets of independent variables, and subtracting the fitted values.

[5] MF acknowledge their strategy likely fails to control for all factors that could bias their estimates. They write (p. 319): "...estimated wage effects of marriage and children may be biased by unmeasured heterogeneity. In other words, women may have self-selected into different marital and fertility states on the basis of unmeasured characteristics that were correlated with earnings." The example I gave above about preferences for work versus family life is a specific example of the type of "unmeasured heterogeneity" MF probably had in mind.

The best I can do here is to say, we should control for those factors that meet the two OVB conditions, except in cases where such a factor is itself an outcome related to the main variable of interest. If we only use control variables that were determined before the treatments (marriage and number of children in the MF study) occurred, we'll usually avoid including bad controls. This is not a complete characterization of the requirements for control variables in the regression control strategy, but it is a good rule of thumb.[6]

Part of the difficulty in deciding which variables should be included and which should be excluded is that it is not clear what is the ideal experiment to measure the effect MF have in mind. Would we recruit a sample of women, say through a reality television show like *The Bachelor*, and randomly select half of them to be married while prohibiting the other half, and track their earnings five years later? Describing an ideal experiment is a way of defining the causal effect we are after.

It is always possible to find ways to improve research studies. MF never claim to estimate the precise magnitude of the causal effect of marriage and children on earnings, and they themselves describe what they did as a "first step." I think the MF study lays very useful groundwork for future research to build from, and as a case study for beginning students, it is a wonderful illustration of both the promise and challenge of the *regression control* technique. I encourage those new to the ACS to read their article and explore my file `script11.R`; perhaps one of the readers of this book will develop a more sophisticated approach to measuring the causal effect of marriage on earnings, maybe using a natural experiment.[7] The topic of the next section is a study that does attempt to exploit a natural experiment, via a *difference-in-differences* approach.

Children and the Business Cycle

Some of us are luckier than others in terms of how we are affected by recessions. Some of my students, for example those in the class of 2019, graduated and found a labor-market desperate for competent workers. The

[6] It is possible to characterize these requirements mathematically. The condition for the coefficient on the main independent variable of interest to have a causal interpretation is called "conditional mean independence." See Stock and Watson (2011), p. 232.

[7] As an example of a natural experiment used to measure the marriage premium for males, Ginther and Zavodny (2001) rely on "shotgun weddings" which they define as marriages that occur right before a baby is born, to separate men into control and treatment groups.

class of 2020, as I write this sentence, is facing a grimmer situation, with the economy suffering from months of lockdown due to the coronavirus pandemic. It feels similar to the Great Recession of 2008, when highly competent students graduated and had to start their careers in a period where jobs were scarce. Some things we have no control over, such as the year we were born, end up having large and permanent impacts on our lives.

Recessions can also derail our best laid family plans. It may seem unlikely that economic factors would impact family planning decisions. However recall in Chapter 3 we saw that in the study by Kuka et al. (2019) the immigration policy Deferred Action for Childhood Arrivals (DACA) was found to have a significant effect on reducing pregnancy among teenage moms. While Kuka et al. (2019) study women at the beginning of their childbearing years, Comolli and Bernardi (2015) study women at the end of theirs, and whether the Great Recession of 2008 caused some women in their late 30s to give up having children all together. This study, titled, "The causal effect of the great recession on childlessness of white American women," is the focus of this section.

To understand the approach Comolli and Bernardi (hereafter CB) take, let's first consider a tale of two sisters, Elsa, born in 1968, and Anna born in 1971. In 2004 Elsa was 36 years old, married, but with no children. Three years later, with her biological clock ticking, she turned 39 in 2007, when the economy was still roaring. She and her husband had high-paying, seemingly secure jobs, and so they decided to have a baby. Elsa's daughter was born on Christmas Day, 2007. By her daughter's first birthday (Christmas, 2008) things had changed dramatically for Elsa and her husband, as they had for the rest of the world, as the Great Recession took its toll. Her husband lost his job, they defaulted on their mortgage, and lost their home.

Elsa's younger sister Anna faced a different set of circumstances as she reached the end of her childbearing years. Watching her sister raise their daughter under financial stress, Anna held off on forming a family of her own, and by her 39th birthday in 2010, had to make a decision; "It is now or never," she told her husband. Faced with few jobs and little prospect of selling their tiny condo in the depressed real estate market, they decide on "never."

This fictional story portrays one-way macroeconomic conditions can affect fertility decisions. It's possible other families, faced with a recession, have the opposite reaction, namely, to have kids because the economy is slow and, to put it one way, there's nothing else to do. Still, it's helpful to

keep this anecdote in mind as we review CB's research design, because at first it seems more complicated than it really is. The reason is, rather than examining just women born in 1968 and 1971, they consider two cohorts of women where cohorts are defined by three years rather than just a single year. The control cohort is women born between 1968 and 1970. This is the older cohort who ended their 30s during normal times. The treatment cohort is younger, born between 1971 and 1973, and reached the end of their childbearing years during a recession. One reason for including three birth years rather than single years in cohort definitions is that, even though the ACS is a large survey, doing so gives them a three-times larger sample size. (An end of chapter question guides the reader through an extension of their analysis where the cohorts are defined by single years of birth.)

Before presenting CB's D-in-D estimate, Fig. 5.2 shows trends in childlessness rates among the treatment and control cohorts, for every year from 2004 to 2010.[8] We see the two lines move (mostly) in parallel until 2010, when the figure for the treatment group jumps up dramatically.[9] The fact that Fig. 5.2 shows the trends in childlessness move in parallel until the Great Recession suggests it is a plausible natural experiment.

[8] CB present a figure like this (Fig. 4 in their paper) but using data going back to 2000, the year of inception of the ACS, to show the *pretrends* moved in parallel prior to the natural experiment (the recession). One might say my version of their figure technically shows only intermediate trends, not pretrends, as the data used to estimate the D-in-D model we will see shortly begins in 2004 and ends in 2010. I didn't include data from 2000 to 2003 in my figure because the master data file for this book doesn't contain ACS samples before 2004. That may not be a good excuse, but I created an exercise, described in the Review Questions to this chapter, that guides a reader through downloading these earlier data and revising Fig. 5.2 to show pretrends.

[9] Why would the childlessness rate increase from 2009 to 2010? Moreover, how could it even increase? Two ways measured childlessness could increase in the population are if children were leaving their parent's homes, or children were dying. Neither of these seems particularly plausible to me. Instead, I think the sample statistics show an increase in childlessness between 2009 and 2010 because of *sampling variation*; the 2009 sample statistic was a little under, and the 2010 statistic was a little above the true population figures. In other words, childlessness probably continued to decrease in the population of women in the 1971–1973 birth cohort from 2009 to 2010, even though it increased in the samples surveyed by the ACS. It is also possible that, although the Census Bureau strives for representative sampling every year, a slightly different population of women was surveyed in 2010 than in prior years. If so, *sample selection bias* could be a factor here as well. For the most part, these are nuances that can easily distract from the thread of the discussion in the text, which focuses on explaining CB's modeling approach.

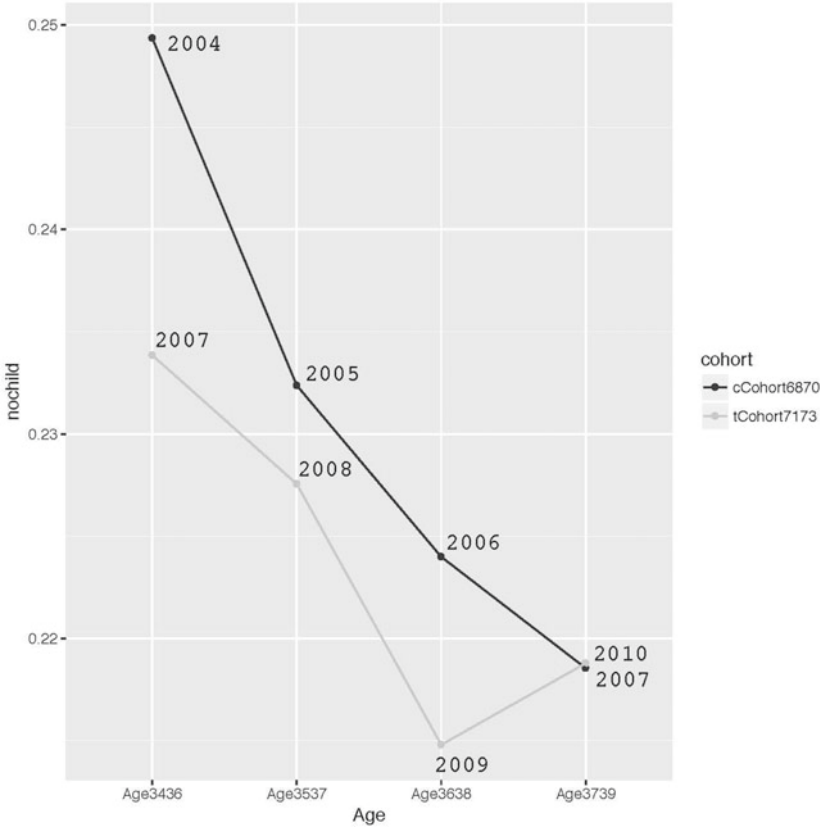


Fig. 5.2 Childlessness and the Great Recession

Figure 5.2 shows childlessness rates for the two cohorts for seven years, but only four of these numbers are needed to calculate the basic D-in-D. These four numbers are shown in Table 5.1. The D-in-D estimate here, like in the previous two chapters, is just the difference between two differences in means. Table 5.1 illustrates this, by presenting the proportion of women in two age groups (centered on 35 and 38, respectively) and from two different birth cohorts (centered on 1969 and 1972, respectively) that report having zero children. This table looks slightly different from the previous tables of D-in-D results; it's not fundamentally different, however,

Table 5.1 Childless rates among women in two birth cohorts

	2004	*2007*	*2010*	*Difference*
Born 1968–1970	24.94	21.85		24.944 − 21.85 = 3.08
Born 1971–1973		23.39	21.88	23.39 − 21.88 = 1.51
Difference-in-differences				3.08 − 1.51 = 1.57

I just added one additional column to make it clearer how birth cohort is used to separate the sample into control and treatment groups. It's easy to confuse the group (control versus treatment) definition, which is year of birth cohort, with the time definition, which is the postrecession period.[10] For the older cohort, women born between 1968 and 1970, 24.94% of the estimation subsample reported zero children in 2004 (when they were aged 34–36), and the figure fell to 21.85% three years later in 2007.[11] Naturally, the rate of childlessness decreases as a cohort has more time to have a child. The reduction is 3.08 percentage points, as shown in Table 5.1.

We now consider women in the younger cohort, born between 1971 and 1973. As these women were nearing the end of their childbearing years, many had to deal with the added stress of a weak labor market and tight credit constraints. The proportion of women in the 1971–1973 cohort without a child was 23.39% when they were surveyed in 2007 at ages 34–36. This is a lower proportion than for the older, control cohort; using data from other sources, CB document that declining fertility is part of a long-term trend in U.S. society. As this cohort aged, the proportion fell to 21.88% when the treatment cohort was surveyed in 2010. This is a reduction of only 1.51 percentage points.

Had the change in childlessness of the younger cohort of women mirrored the older cohort, it would have fallen by 3.08 percentage points. Thus, the younger cohort would have had a childlessness rate of 23.39 −3.08 = 20.31%, if not for the Great Recession. Instead, the rate was

[10] The numbers in Table 5.1 are those my student and I found in our replication of the CB study. Most of our numbers match exactly with those from the original study. The number of observations in the estimation subsample, and two of the coefficient estimates from the basic D-in-D regression model are identical, but, the two other coefficient estimates are slightly different.

[11] It's worthwhile to note, the same women weren't surveyed in 2004 and 2007. The ACS is a repeated cross-section not a panel survey.

21.88, a difference of 21.88 − 20.31 = 1.57. This figure, 1.57 percentage points, is the D-in-D.

As we've seen in the previous two chapters, this D-in-D estimate can also be found by subtracting one difference in means from another. Previously I've shown this in the bottom row of the tables, and I continue to do this in Table 5.1. The averages reported in Table 5.1 can be estimated with a multivariate regression model with two binary variables and an interaction term. The form of the equation is shown below:

$$NOCHILD_{it} = \alpha_0 + \alpha_1 COHORT7173_i + \alpha_2 AGE3739_t$$
$$+\alpha_3 (COHORT7173_i \times AGE3739_t) + u_{it},$$

where the dependent variable $NOCHILD_i$ is equal to one if the woman does not have a child at home, and it is equal to zero if they have one or more children, $AGE3739_t$ is a binary equal to one if the woman is between the ages of 37–39 and is zero if they are between 34 and 36, and the variable $COHORT7173_i$ is a binary variable equal to one if the woman was born between 1971 and 1973 and is equal to zero if they were born between 1968 and 1970.

Estimating this equation on a subsample of white women who do not live in group quarters, we find the estimated regression coefficients shown below:

$$NOCHILD_{it} = 0.249 - 0.016 \times COHORT7173_i - 0.031 \times AGE3739_t$$
$$+0.0157 \times (COHORT7173_i \times AGE3739_i).$$

To interpret this equation, try plugging in various combinations of the two independent variables and convince yourself that the interaction regression equation produces predictions that are equal to the sample proportions reported in Table 5.1. For example, say both $AGE3739_i$ and $COHORT7173_i$ are equal to zero. Then the equation reduces to 0.249, which is the childlessness rate of the control (older) cohort when they were aged 34–36.

Like the previous two case studies from Chapters 3 and 4, CB estimate more elaborate variants of the basic D-in-D model. The fact that results from more elaborate models agree with results from the basic D-in-D bolsters confidence in their findings. The specific variants of the basic D-in-D model they estimate is another example of a *basic D-in-D with control variables*, which we first saw in Chapter 3. In one of their models, they add

state fixed effects only to the basic D-in-D model, the equation for which is shown below:

$$NOCHILD_{its} = \beta_0 + \beta_1 COHORT7173_i + \beta_2 AGE3739_t$$

$$+\beta_3(COHORT7173_i \times AGE3739_t) + \sum_{k=Alaska}^{Wyoming} \beta_k STATE_{k_s} + \varepsilon_{its},$$

The addition of state fixed effects to the basic D-in-D model offers a way of controlling for factors that vary across states but are unchanged over time, such as climate, cultural attitudes, and so on. The beauty of using fixed effects to control for these factors is we do not have to actually measure them and can instead rely on binary variables. The estimate of β_3 is 0.0163. This is very close to the value of 0.0157 we find in the basic D-in-D without controls.

The Comolli and Bernardi study is about fertility, and hence why I discuss it in this chapter on family economics. However, their basic approach can be used to study the effect of the Great Recession on other outcomes. For example, Coile and Levine (2011) study retirement, and the impact on earnings of retiring in a recession versus a boom time. Both of these studies use birth cohorts to define the control and treatment group. These studies analyze different outcomes, but are united by their use of recessions as natural experiments. In the next chapter, we'll see an example of a study that unites the subdisciplines of family economics and transportation economics.

Key Terms

Regresson control	Difference-in-differences (D-in-D)	Polynomial model
Bad Control	Post-treatment variable	Pretrends
Sampling variation	Sample selection bias	Basic D-in-D with control variables
Endogeneity		

Questions for Review

1. Review Table 1 of the study by Marshall and Flaig (2014) which carries out a difference in means test, and use it to discuss logarithmic data transformations (a concept we first saw in Chapter 2). Then, review their Table 3 and use it to discuss polynomial data transformations.
2. Table 5.2 contains data from six households that are comprised of one mom, one dad, and either one or two children. Use these data to carry out a difference in means test of the null hypothesis that moms with one child work the same number of hours as moms with two kids, using a paper and pencil. To do this without a calculator it helps to know that the square root of 361 is 19 and the square root of 400 is 20. After trying it by hand, write an R script to automate analysis of these data, by calculating summary statistics, tabulating the data, creating a scatter plot, and carrying out a difference in means test in two ways: with a t-test and with a regression (Hint: Start with `script4.R` and modify it as needed). Finally, discuss whether the regression estimate suffers from selection bias.
3. The study by Marshall and Flaig (2014), discussed in this chapter presents regression models for self-employment earnings on marriage and children, but the authors note, "This analysis may have suffered from...endogeneity...marital status and number of children may be endogenous...women may have self-selected into different marital and fertility states on the basis of unmeasured characteristics that were correlated with earnings" (p. 319). Omitted variable bias is one form of *endogeneity*. Give an example of an omitted variable that both predicts earnings and is correlated with marriage status or number of children (Hint: preferences for careers versus family life). Can you think of an ideal experiment to study the effect of children on earnings? (Hint: orphanages). Can you think of a natural experiment to study the effect of marriage and children on earnings? (Hint: shotgun weddings).
4. This question is adapted from a Discussion Question in Sect. 8.5 of Michael Bailey's textbook, *Real Econometrics*. He cites Rossin-Slater et al. (2013). Explain how to create a simple D-in-D estimate of the effect of the policy for the following example: California implemented a first-in-the-nation program of paid family leave in 2004, and we wish to know if this policy increased the use of maternity leave. The outcome of interest here is the proportion of new mothers

Table 5.2 ACS Data, 2015, PUMA 068511, select households

SERIAL	*Number kids*	*Mom's work hours*
68781	2	0
69519	2	45
71181	2	0
71283	1	35
73412	1	0
81644	1	40

Notes The end of the Data section of the file `script1.R` produces a CSV file with data from these households. Data from household 68781, shown here, was also discussed in Chapter 1, and appeared in Tables 1.1 and 1.2

(women having a baby in the last six months) who are working. Use the basic D-in-D regression equation shown below:

$$Y_{it} = \beta_0 + \beta_1 TREAT_i + \beta_2 POST_t + \beta_3 (TREAT_i \times POST_t) + u_i,$$

and explain how you would construct each of the three variables: Y_i, $TREAT_i$ and $POST_t$. Finally, describe the estimation subsample.

5. The Comolli and Bernardi (2015) study described in this chapter uses an estimation subsample of white women. Modify the file `script12.R` and extend the analysis by estimating their D-in-D models for a different population of women.
6. The `ACSmaster.RData` file described in Appendix A contains ACS samples from 2004 to 2017. Download ACS data from IPUMS from 2000 to 2010, and modify the file `script12.R` to calculate pretends, as in Comolli and Bernardi's (2015), Fig. 4.
7. Modify the file `script12.R` so cohorts are defined based on one year of birth rather than three as in the case study here. How does the definition of control and treatment groups affect the results?

References

Abraham, Katharine G., John Haltiwanger, Kristin Sandusky, and James Spletzer. "The rise of the gig economy: Fact or fiction?" AEA Papers and Proceedings 109 (2019): 357–361.

Angrist, Joshua D., and Jörn-Steffen Pischke. Mastering 'metrics: The path from cause to effect. Princeton University Press, 2014.

Argys, Laura M., and Susan L. Averett. "The effect of family size on education: New evidence from China's one-child policy." Journal of Demographic Economics 85, no. 1 (2019): 21–42.

Bailey, Michael A. Real econometrics: The right tools to answer important questions. Oxford University Press, 2017.

Bailey, James, and Dhaval Dave. "The effect of the Affordable Care Act on entrepreneurship among older adults." Eastern Economic Journal 45, no. 1 (2019): 141–159.

Blau, F. D., L. M. Kahn, P. Brummund, J. Cook, and M. Larson-Koester. "Is there still Son preference in the United States?" Journal of Population Economics 33, no. 3 (2020): 709–750.

Coile, Courtney C., and Phillip B. Levine. "Recessions, retirement, and social security." American Economic Review 101, no.3 (2011): 23–28.

Comolli, Chiara Ludovica, and Fabrizio Bernardi. "The causal effect of the great recession on childlessness of white American women." IZA Journal of Labor Economics 4, no. 1 (2015): 1–24.

Costa, Dora L., and Matthew E. Kahn. "Electricity consumption and durable housing: Understanding cohort effects." American Economic Review: Papers & Proceedings 101, no. 3 (2011): 88–92.

Dahl, Gordon B., and Enrico Moretti. "The demand for sons." The Review of Economic Studies 75, no. 4 (2008): 1085–1120.

Ginther, D. K., and M. Zavodny. "Is the male marriage premium due to selection? The effect of shotgun weddings on the return to marriage." Journal of Population Economics 14, no. 2 (2001): 313–328.

Kuka, Elira, Na'ama Shenhav, and Kevin Shih. "A reason to wait: The effect of legal status on teen pregnancy." AEA Papers and Proceedings 109 (2019): 213–217.

Marshall, M. I., and A. Flaig. "Marriage, children, and self-employment earnings: An analysis of self-employed women in the US." Journal of Family and Economic Issues 35, no. 3 (2014): 313–322.

Munger, M. C. Tomorrow 3.0: Transaction costs and the sharing economy. Cambridge University Press, 2018.

Rossin-Slater, Maya, Christopher J. Ruhm, and Jane Waldfogel. "The effects of California's paid family leave program on mothers? Leave? Taking and subsequent labor market outcomes." Journal of Policy Analysis and Management 32, no. 2 (2013): 224–245.

Schneebaum, Alyssa, and M. V. Lee Badgett. "Poverty in US lesbian and gay couple households." Feminist Economics 25, no. 1 (2019): 1–30.
Stock, James H., and Mark W. Watson. Introduction to econometrics. Boston, MA: Addison Wesley, 2011.

PART IV

Instrumental Variables

Learning Goals for Part IV

1. Explain why a regression of vehicles on neighborhood population density suffers from self-selection bias.
2. Compare and contrast the method of instrumental variables (IV) with regression control and D-in-D.
3. List the two conditions for a valid instrument.
4. Assess whether sibling gender is a valid instrument in the population density-vehicle demand example in this chapter.

CHAPTER 6

Getting Around: Cars and Land Use

The ACS asks Americans several questions about their transportation behavior. With the ACS we know how many vehicles a household has, for each worker we know how long their journey to work takes, and their mode of commuting (driving, transit, bicycle, etc.). It is also possible to estimate the distance between a worker's home and work locations. The Covid-19 pandemic put into sharp focus the importance of working from home. From the ACS we learn 41% of those working as writers worked from home in 2017, and they were likely well-suited to the stay-at-home orders most Americans found themselves under in Spring of 2020. Other industries and occupations were more affected; zero respondents in some manufacturing industries reported working from home.[1] There is a lot we can learn about transportation behavior from the ACS.[2]

Transportation is important because in a typical recent year (say, 2017) the average American worker spent 26.6 minutes commuting one way. A

[1] Some manufacturing, like operating industrial scale machines, cannot be done in the home, but during the early days of the pandemic I watched my childhood friend on Facebook move some of his manufacturing to his garage, so he could help care for his young children. For more details on this analysis, see my blog article at: www.mattholian.blogspot.com/2020/04/where-do-most-people-work-from-home.

[2] There are surveys specifically designed to measure transportation behavior. The National Household Travel Survey (NHTS), for example, is much more detailed than the ACS in this regard. However, it surveys far fewer people.

M. J. Holian, *Data and the American Dream*,
https://doi.org/10.1007/978-3-030-64262-4_6

worker who spends one hour in their daily commute, five times a day, fifty weeks a year, is spending 250 hours, or more than ten straight days per year commuting. Traffic accidents are a leading cause of death for Americans, and in addition transportation is responsible for about 15% of the average U.S. household's carbon emissions (Nordhaus 2013, p. 161).

We saw in Chapter 2, in the context of the Costa and Kahn (2011) case study, that home energy efficiency regulations represent one approach to lowering household carbon footprints. In the area of transportation, energy efficiency regulations are referred to as "CAFE standards" (Corporate Average Fuel Economy) in the US, and require vehicle manufacturers to achieve increasingly higher levels of fuel efficiency as measured by miles-per-gallon. While this by itself is certainly good for the environment and worth celebrating, as we saw in the case of homes, by lowering the cost of using energy, energy efficiency regulations can also have the unintended effect of encouraging households to drive more (the so-called "rebound effect," which implies that the benefits of CAFE standards from the environment could be at least partially offset by increases in traffic accidents).

A different approach to reducing energy use from transportation involves "land-use policies"—getting people to drive less by encouraging households to live in urban areas where they can take transit and walk. Land-use policies local governments enacted to raise density include, as two examples, Minneapolis which recently changed its zoning to allow the conversion of single-family homes in suburban neighborhoods into multi-family housing, and San Francisco, which began permitting the construction of tall buildings after a long period of resisting "Manhattanization." Transportation plays a key part in our lives, and it connects to several themes from earlier chapters, including gentrification, energy regulation, ride-sharing, and the gig economy. The pandemic certainly exposed how workers in some occupations are better able to avoid travel than others; nonessential professionals in urban areas worked from home while gig workers delivered groceries and takeout.

The next section of this chapter examines the connection between land-use and transportation.

Car Use in Compact Versus Sprawling Cities

Does living in a high-density neighborhood really cause people to drive less, or do people who dislike driving choose to live in dense neighborhoods?

This is the key question we study in this section. Figure 6.1 shows a suggestive scatterplot, with the number of vehicles a household has access to on the vertical access, and the population density of the PUMA in which the household lives on the horizontal. Figure 6.1 also fits a regression line that is downward sloping; households in higher density areas have fewer vehicles. The regression equation here is a bivariate model with continuous dependent and independent variables:

$$VEHICLES_j = \gamma_0 + \gamma_1 DENSITY_j + \varepsilon_i,$$

where $VEHICLES_j$ is equal to the number of vehicles household j has access to (we don't know if they are owned, leased, borrowed, etc.) and the independent variable $DENSITY_j$ is measured as the number of people in the PUMA in which household j lives, divided by its land area.[3] PUMAs in areas like New York City have the highest population density in the country, while those in the Nevada desert are the lowest. Suburban areas, where most Americans live, are in between, and vary in terms of density. Some suburbs have homes fairly close together, with shops and transit lines close enough to walk to, while in other suburbs, beyond going for a walk, residents are essentially dependent on a car for all trips.

Even without seeing Fig. 6.1, we would expect the coefficient on γ_1 to be negative, simply because our experience tells us households need fewer vehicles in urban areas and vice versa. The estimated equation is shown below:

$$VEHICLES_j = 2.34 - 0.00003 \times DENSITY_j + \varepsilon_i,$$

The negative coefficient on $DENSITY_j$ in the equation above is seen as the negative slope of the line in Fig. 6.1. In high-density areas like Manhattan people own few and perhaps zero cars, while in low-density areas some households own six or more vehicles.[4] It just so happens that not a single household in PUMAs with more than 45,000 people per square mile owns more than three vehicles in this sample of 379,117.

Although vehicle ownership falls as density increases, the connection isn't necessarily causal in the sense that Fig. 6.1 doesn't prove that making

[3] In Chapter 1, Table 1.3 showed land area for select PUMAs. PUMAs are drawn to have around 100,000 people in them, but they vary considerably in land area.

[4] Households with more than six cars are listed as having six; this is an example of *top-coding* the data, which helps preserve the anonymity of respondents and prevents errors in reporting or recording from skewing estimates.

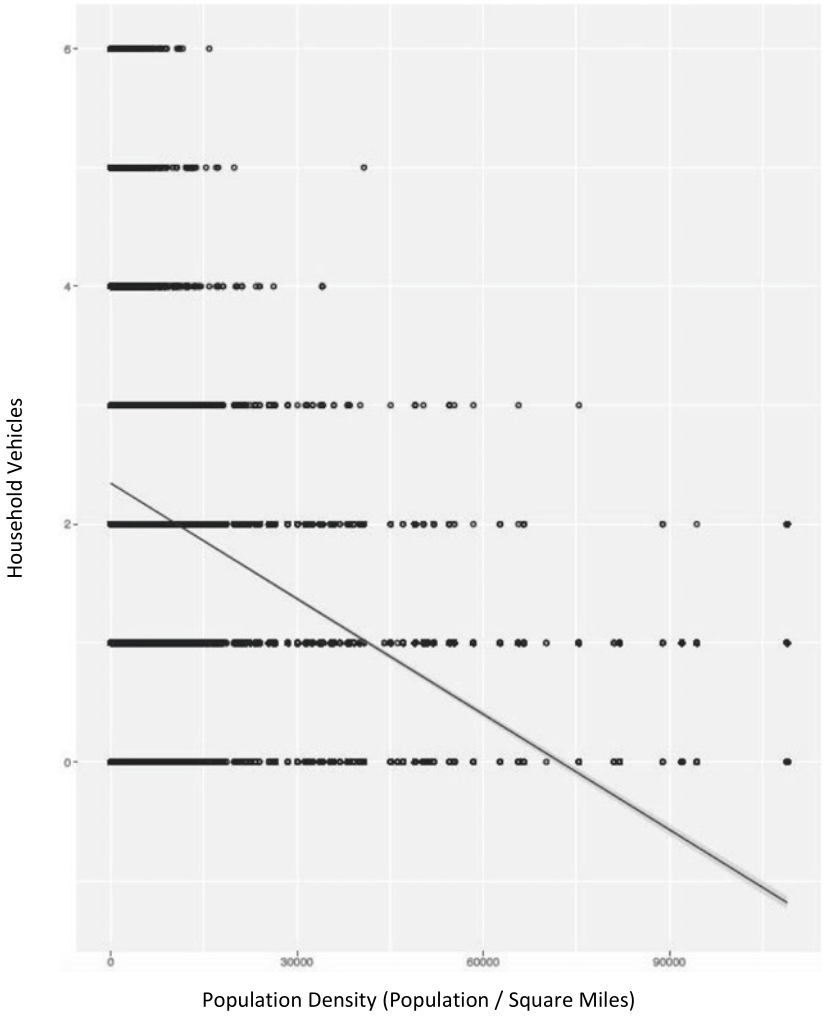


Fig. 6.1 Household vehicle ownership and population density. The sample consists of married-couple households with exactly two children where the head of household is white and between 25 and 55. ACS samples 2012–2017

neighborhoods denser will cause people to own fewer vehicles and to drive less. Maybe people who already hate driving choose to live in dense areas, and people who love cars and driving choose to live in low-density suburbs. If this is the case, then if we moved someone who loves driving from a low density to a high-density location (an *ideal experiment*), their transportation decisions may not change much. At the very least, we would expect that Fig. 6.1 overstates the magnitude of the causal effect of dense urban living on driving due to this selection effect.

In principle, we could use regression control to eliminate bias from these other factors, just as we did in Chapter 2 to study home energy codes (where we included variables such as number of rooms that are correlated both with the energy expenditure and home vintage), and in Chapter 5 to study the effect of marriage on labor-market outcomes (where we included variables such as age, that are correlated with earnings and probability of being married). In the case of estimating the causal connection between neighborhood density and vehicle ownership, we would want to include a control variable that measures the household's preference for driving, as this is correlated with both neighborhood choice and vehicle ownership.[5] It is possible to think of some variables that proxy for driving preferences, but the ACS does not really contain any compelling measures of it. Surveys like the ACS simply don't ask questions about whether the respondents think public buses are "icky," driving is "scary" or other topics that would allow us to control for travel preferences.

In a study titled, "An Empirical Analysis of Urban Form, Transport, and Global Warming," Fabio Grazi and his coauthors Jeroen C.J.M. van den Bergh and Jos N. van Ommeren, study the causal effect of density on commuting distance and mode, using Dutch housing survey data and an empirical approach called *instrumental variables* regression, or IV for short.[6] IV is a more advanced technique than I have discussed in this book until now. I include a discussion of it because the Grazi et al. (2008) study illustrates the IV technique especially clearly. The key to IV regression is to

[5] Now is a good time to review the two OVB conditions (see Chapter 1 or the glossary) because in each of these examples, a control variable is included that meets the two OVB conditions to reduce bias in the estimate of the coefficient on the main independent variable of interest.

[6] So far in this book I have been discussing replications and extensions of studies that originally used the ACS (or its precursor, the long-form decennial Census), but here we see an example where the ACS can be used to estimate models that were initially developed with data from entirely different countries.

think of a compelling *instrument*. In this case, a valid instrument would be a variable that predicts a household's neighborhood density, but does not affect vehicle ownership directly, except through its influence on density.[7]

Our transportation, housing, and labor-market decisions are obviously influenced by our family situations, but sometimes in surprising ways. We saw in Chapter 1 that families where the children have the same gender live, on average, in homes with slightly fewer bedrooms. Using a sample of married couple households with exactly two children, I estimated a bivariate regression equation with number of $BEDROOMS_j$ in household j on the left-hand side and $SAMEGENDER_j$ (a binary variable equal to one if the children in the household are either both boys or both girls) on the right. $SAMEGENDER_j$ is an instrumental variable proposed by Grazi et al. (2008). As I reported in Chapter 1, I find households with same gender children have homes with slightly fewer bedrooms, on average, than households with different gender children. The estimated coefficient on $SAMEGENDER_j$ is negative and statistically significant, though not very large in magnitude.

Why does this matter? If families with same gender children need fewer rooms, then these families will have more options in dense neighborhoods where homes are typically smaller and that are usually more walkable and transit friendly. If families with same gender children live in denser neighborhoods, then $SAMEGENDER_j$ is a *relevant instrument*. To test for relevance, I use ACS data from 2012 to 2017 to estimate the following equation:

$$lnDENSITY_j = \alpha_0 + \alpha_1 SAMEGENDER_j + \alpha_2 lnHHINCOME_j$$

$$+ \alpha_3 COLLEGE_j + \alpha_3 WORKERS_j + \varepsilon_i,$$

where the dependent variable $lnDENSITY_j$ is the (natural log of) population density in the PUMA in which household j lives, and our main independent variable of interest is $SAMEGENDER_j$, the binary variable described above that is equal to one if both children in the household have the same gender (recall our sample includes only two-child households). The rest of the right-hand side variables are control variables; $COLLEGE_j$ is a

[7] The two conditions are: valid instruments must be exogenous (they do not directly affect the dependent variable) and relevant (it must be a statistically significant predictor of the main independent variable of interest). See Angrist and Pischke (2014, Chapter 3), and Bailey (2017, Chapter 9), or Stock and Watson (2011, Chapter 12).

binary variable equal to one if the head of household has a college degree, $WORKERS_j$ is equal to the number of people in the household that work, and $lnHHINCOME_j$ is (the natural log of) household income.

Rather than report all of the estimated coefficients in this equation (which can be found in the article this discussion is based off of, Holian 2020), I report just the key coefficient: I find the estimate of α_1 to be equal to 0.02. Like the effect of $SAMEGENDER_j$ on a household's bedrooms, the effect of child gender on neighborhood log density is not large in magnitude. But given our sample size is large and the households in our estimation subsample so comparable, this small effect is statistically significant. We predict a married couple household with two kids with same gender children will live in a slightly more compact neighborhood than a household with a boy and a girl. This establishes that $SAMEGENDER_j$ is a relevant instrument, the first of the two conditions for valid instruments.[8]

The second condition for $SAMEGENDER_j$ to be a valid instrument is for it to be an *exogenous instrument*. Here this means the gender of the children cannot directly affect the number of vehicles in a household. I admit it's possible to imagine that families where the children have the same gender might economize on some trips; for example, taking both girls to ballet class or both boys to football, and needing fewer vehicles on average as a result. It's not possible to test this with the ACS, but because, as argued in the last chapter, the gender of children is as good as randomly assigned, it does not seem likely that it is correlated with underlying travel preferences of the parents. Thus, I argue $SAMEGENDER_j$ is also an exogenous instrument.[9]

The IV technique works by estimating two equations sequentially. The *first-stage equation* was the equation above where $lnDENSITY_j$ was the dependent variable, and the *second-stage equation* is below:

$$VEHICLES_j = \beta_0 + \beta_1 ln\hat{DENSITY}_j + \beta_2 lnHHINCOME_j$$

[8] In this case, the test boils down to whether or not the value of the test statistic, in a test of the hypothesis that the coefficient on $SAMEGENDER_j$ is zero, is greater or less than 3.33. See Stock and Watson (2011, p. 439) who describe a rule of thumb for checking instrument relevance.

[9] Unlike the relevance condition, there is no formal test for the exogeneity condition. Instead, a plausible argument, or assumption, is generally required to establish it. It would be possible in principle to bring evidence to bear that reinforces my assumption, for example, I could use data from the NHTS to see if families with same gender children have same or different number of trips, mileage, or gasoline expenditures.

$$+ \beta_3 COLLEGE_j + \beta_3 WORKERS_j + \varepsilon_i.$$

In this second-stage equation, instead of using the household's actual log density on the right-hand side, we use the predicted value of the household's density from the first-stage equation.[10] The hat on $\hat{lnDENSITY}_i$ in the equation above denotes that it is the predicted value of density for household j, not the actual density of the PUMA in which they live. The actual density they live in is determined by many factors including their preference for travel, but the value of density predicted by the gender of their children should not be correlated with their preferences for travel, which is what allows the IV technique to solve selection bias.

The results of estimating this second equation using the two-stage IV model are shown below:

$$VEHICLES_j = 0.34 - 0.12 \times \hat{lnDENSITY}_j + 0.19 \times lnHHINCOME_j$$

$$- 0.08 \times COLLEGE_j + 0.28 \times WORKERS_j + \varepsilon_i.$$

These results indicate that the causal relationship between density and vehicle ownership may not be very different from the simple correlation between these variables. In other words, the line in Fig. 6.1 might not be that biased after all.[11]

Commuting is a major source of household carbon dioxide emissions. The next chapter shows how to incorporate climate change considerations into a formal Cost–Benefit Analysis of energy policy. This concluding chapter considers how to evaluate policy options, and asks the question, "What should we do?" Making correct policy choices in a complicated world is not easy. Luckily, CBA can help guide us to efficient decisions.

[10] To do this, we plug in the values of the independent variables for all households into the estimated first-stage equation, to calculate *fitted values*. These can also be called predicted values.

[11] Understanding how I come to this conclusion requires some unpacking. Technically, Fig. 6.1 plots density while the IV equation uses log density as a dependent variable, so the estimates aren't actually directly comparable. In Holian (2020), I estimate a version of the second-stage equation using actual $lnDENSITY_j$ rather than predicted $\hat{lnDENSITY}_j$ which amounts to a regression control approach but might produce biased estimates because it omits a variable that measures travel preferences. The estimated coefficient on $lnDENSITY_j$ is -0.09 while it is -0.12 in the IV model. To me this is a small difference, it suggests the direction of bias is not as I expected, and so I conclude the simple, bivariate regression (the non-IV approach) may not be very far off from the causal effect we are after.

Key Terms

Top code	Ideal experiment	Instrumental variables (IV)
Instrument	Valid instrument	Relevant instrument
Exogenous instrument	First-stage equation	Second-stage equation
Fitted values		

Questions for Review

1. Explain why a regression of vehicles on neighborhood population density suffers from self-selection bias.
2. Compare and contrast the method of instrumental variables (IV) with regression control and D-in-D.
3. List the two conditions for a valid instrument. Is there a statistical test to determine whether an instrument meets both conditions?
4. Assess whether sibling gender is valid instrument in the land-use density-vehicle demand example in this chapter.
5. In the case study in this chapter, the dependent variable was the number of vehicles that a household has access to. The ACS also contains other measures of travel behavior. Can you think of a new way to measure travel behavior with the ACS, modify `script13.R` and carry out original research?

References

Angrist, Joshua D., and Jörn-Steffen Pischke. Mastering 'metrics: The path from cause to effect. Princeton University Press, 2014.

Bailey, Michael A. Real econometrics: The right tools to answer important questions. Oxford University Press, 2017.

Costa, Dora L., and Matthew E. Kahn. "Electricity consumption and durable housing: Understanding cohort effects." American Economic Review: Papers & Proceedings 101, no. 3 (2011): 88–92.

Grazi, F., J. C. van den Bergh, and J. N. van Ommeren. "An empirical analysis of urban form, transport, and global warming." The Energy Journal 29, no. 4 (2008): 97–123.

Holian, Matthew J. "The impact of urban form on vehicle ownership." Economics Letters 186 (2020): 108763.

Nordhaus, William D. The climate casino: Risk, uncertainty, and economics for a warming world. Yale University Press, 2013.

Stock, James H., and Mark W. Watson. Introduction to econometrics. Boston, MA: Addison Wesley, 2011.

PART V

Putting Estimates Into Action: Econometrics and Cost–Benefit Analysis

Learning Goals for Part V

1. Define Cost-Benefit Analysis (CBA), and describe a difference between CBA and other forms of economic analysis.
2. List the steps of a CBA and identify steps where accurate descriptive statistics and causal inference are especially important.
3. Compare and contrast the do-it-yourself and plug-in methods for estimating and monetizing impacts.

CHAPTER 7

Conclusion: What Do We Know and What Should We Do?

The preceding chapters have illustrated econometric methods for empirically estimating both accurate descriptions of reality and causal effects by presenting examples of research that has used the American Community Survey. The focus of this final chapter is rather different. The key question I address here is, how should these empirical estimates be used to guide public policy decision-making?

Many of the research studies we saw in earlier chapters directly relate to specific areas of public policy, such as: building energy codes, immigration, college financial aid, occupational licensing, public health insurance, and many more. The subfield of economics that studies public policy decision-making is called *Cost–Benefit Analysis* (CBA). This chapter describes CBA, and how it crucially depends on having credible empirical measures. By way of example, this chapter walks the reader through a case study of assessing building energy code changes in the state of Florida. Building energy codes help us save energy resources, slow climate change, and reduce air pollution, but they also make building homes more expensive. This makes life harder for families and may exacerbate problems like gentrification. Are they worth it? CBA is a tool for answering precisely this sort of question.

The title of this final chapter is, "What do we know and what should we do?" Careful empirical research can help us determine *what we know* about cause and effect relationships in public policy. There is obviously a great deal of disagreement in public discourse today about basic facts; one of my hopes is that better empirical research will help make establishing facts less

M. J. Holian, *Data and the American Dream*,
https://doi.org/10.1007/978-3-030-64262-4_7

controversial.[1] However, even when there is agreement on facts, there can still be disagreement about what we should do in terms of public policy. My answer to the question *what should we do?* is we should, whenever prudent and to the extent it is practical, use CBA. I don't claim CBA will be able to answer all questions, and I don't have answers to all of the policy questions raised in this book, but I do have a suggestion for how society should, to a greater extent than we do now, go about deciding them.

My proposal to rely more on CBA to determine what we should do in the public policy arena is similar in some ways to the answer Benjamin Franklin gave his friend Joseph Priestley, in 1772, when he asked Franklin for advise on whether or not he should accept a tempting job offer. Here was Franklin's response[2]:

> Dear Sir,
> In the Affair of so much Importance to you, wherein you ask my Advice, I cannot …advise you what to determine, but if you please I will tell you how. …my Way is, to divide half a Sheet of Paper by a Line into two Columns, writing over the one Pro, and over the other Con. Then …I put down under the different Heads …Motives …for or against the Measure. When I have thus got them all together in one View, I endeavour to estimate their respective Weights; and where I find two, one on each side, that seem equal, I strike them both out: If I find a Reason pro equal to some two Reasons con, I strike out the three. If I judge some two Reasons con equal to some three Reasons pro, I strike out the five; and thus proceeding …I come to a Determination accordingly. And tho' the Weight of Reasons cannot be taken with the Precision of Algebraic Quantities, yet when each is thus considered separately and comparatively, and the whole lies before me, I think I can judge better, and am less likely to make a rash Step; and in fact I have found great Advantage from this kind of Equation, in what may be called Moral or Prudential Algebra.
> B. Franklin

[1] There's no better motivation for this than the Covid-19 pandemic. Were the lockdowns instituted in most states starting in March of 2020 good public policy? Answering this requires knowledge of their causal impacts, as well as making philosophical judgments about tradeoffs like *the value of a statistical life*. The pandemic clearly illustrates the strong social need for people who understand how to estimate the impacts of policy, and who also understand techniques for rational decision-making.

[2] https://founders.archives.gov/documents/Franklin/01-19-02-0200.

Although Franklin was advising his friend to use "moral algebra" to make the purely individual decision of whether or not to accept a job offer, economists have devised the related framework of CBA to guide public decisions. Rather than focusing on what is best for the individual, CBA focuses on what is best for society. Society can be defined various ways, as we will see, but in all cases it refers to some group of individuals. Contemporaries of Franklin, most notably Jeremy Bentham, argued that all individuals in society should be treated equally—a radical idea in places across the globe at the time—and this notion forms the basis of how CBA is conventionally practiced today. An objective of, "...the greatest happiness of the greatest number," which is how Bentham put it, elevates the common-man to an equal footing with kings and queens. Another difference is the CBA framework measures all "reasons" or "motives" in dollar values, and this provides a more systematic way of comparing benefits and costs than moral algebra's "strike out the three" costs for one benefit way.

There are entire textbooks on CBA and in this one chapter I won't do justice to all the technical complications and philosophical nuances.[3] As in the preceding chapters on econometric methods, I illustrate the technique of CBA mainly by example. Reproducing and describing calculations that appear in published research reveals a lot about what CBA is—and what it is not.

While I advocate for greater use of CBA, both by public decision makers and researchers, I don't believe it is a panacea for all of society's controversies. Like all research methods, CBA has its drawbacks, and I discuss some of these in the final section of this chapter. CBA also requires analyst judgment, so while it's safe to say that economists of all stripes support its use, two different analysts can use CBA to study the same question and arrive at different conclusions. My hope is the following case study will illustrate the potential of CBA as well as some of the important considerations when using it.

Cost–Benefit Analysis and Energy Codes, Revisited

We saw in Chapter 2, in our review of the Costa and Kahn (2011) study, that building energy codes are public policies aimed at reducing energy consumption. First enacted by some U.S. states in the late 1970s, they

[3] See Boardman et al. (2017) and Fugiutt and Wilcox (1999).

have become more widespread and stringent over time. In Chapter 2 we asked, How much did a state's initial adoption of energy codes cause energy use to fall? In this chapter, we ask, Given what we know about the impact of energy codes on energy use, is making energy codes stronger a wise policy from the standpoint of society broadly conceived?

This section describes a recent *economic analysis* of changes to building energy codes in Florida, which was carried out by Grant Jacobsen and Matthew Kotchen (2013), hereafter JK, in a study titled, "Are building codes effective at saving energy? Evidence from residential billing data in Florida." Although strictly speaking JK calculate what I would call social payback periods for a representative household, their analysis shares a lot in common with CBA. The file `script14.R` carries out the CBA described below, which is my adaptation of the economic analysis presented in the JK study.[4]

CBAs are typically carried out in one of two settings. First, in the course of complying with various mandates, government agencies may commission CBAs or carry them out themselves. These studies typically strive to be comprehensive and adhere closely to the principles of CBA, but the quality of government CBAs varies. Second, academic journals sometimes publish CBAs, but more often it is a component to a larger study rather than the authors' exclusive focus. This is exemplified by the JK study. Its focus was empirically estimating the impact of the energy codes on energy demand—this is the same general question Costa and Kahn (2011) studied—and JK carried out an economic analysis as a secondary part of their study. In terms of length, the economic analysis in JK amounted to just six paragraphs out of a 16-page article. The brevity of their analysis is a virtue for our purpose of illustration.

Most CBAs share a common set of general features. The leading CBA textbook (Boardman et al. 2017, p. 6) describes them in a widely cited list containing nine steps. I reproduce these in Table 7.1.

[4] Other methods of economic analysis include *Economic Impact Analysis* (EIA) and *Fiscal Impact Analysis* (FIA), which are often mistakenly described as CBAs. One of the goals is to section is to describe what CBA is, so a reader will be able to recognize when an analysis that is described as a CBA is in fact something else. The focus in CBA is on human welfare broadly conceived, while EIA and FIA are narrower and focus on specific impacts. For example, FIA may focus on the impact of some policy or program on the state government's budget, while EIA may focus on GDP impacts. Meanwhile, CBA recognizes that social welfare can go up, even as state budgets and GDP go down. Examples of EIA and FIA, respectively, include Chaudhuri and Zieff (2015) and Culhane et al. (2002).

Table 7.1 Nine steps to CBA

1. "Specify the set of alternative projects"
2. "Decide whose benefits and costs count (standing)"
3. "Catalogue the impacts and select measurement indicators"
4. "Predict the impacts quantitatively over the life of the project"
5. "Monetize (attach dollar values to) all impacts"
6. "Discount benefits and costs to obtain present values"
7. "Compute the net present value of each alternative"
8. "Perform sensitivity analysis"
9. "Make a recommendation"

Source Boardman et al. (2017, p. 6)

This list, or minor variations on it, is widely used in the literature. For example, in the context of CBA of crime, Manning et al. (2016, p. 36) describe an essentially identical list that has ten steps. My own opinion is that steps 6 and 7 could be combined together, making this a list of eight steps. Other minor variations have also appeared elsewhere, and I use this list to organize the discussion that follows.

JK examine a change to Florida's energy codes. Florida initially adopted energy codes in 1978 and strengthened them in 2002. The details of Florida's 2002 energy code change are complicated, but as a simplification they frame the change as requiring new homes to use more expensive windows with a low-emissivity (low-E) coating, which should reduce electricity and natural gas demand. This was expected to lower household energy bills.

In terms of CBA Step 1, policy makers in 2002 may have considered multiple alternatives, but this analysis assumes there were two: change the code to require low-E windows, or don't change the code. Most likely, policy makers at the time also considered stronger or weaker versions of the change, or other completely different policy instruments to promote energy efficiency such as taxes or cap and trade, and these other alternatives could in principle be included in a CBA. Like most retrospective analyses, we proceed with a set of alternatives equal to two.

In CBA step 2, *standing* refers to who is considered a member of society. This is a deeply philosophical question, but it is usually decided in CBAs on the basis of practical considerations. For example, a CBA conducted by a government agency may count costs and benefits to U.S. citizens only. However, some economists hold that all impacted parties should have standing (Fuguitt and Wilcox 1999, p. 53). As we see next, the JK analysis

incorporates several distinct definitions of society. In one case they consider a definition where only the homeowner has standing, and another that could be described as one where only citizens have standing. In a third they incorporate climate change impacts that result from the burning of fossil fuels required to produce energy, and given these damages are global in scale, the implicit delineation of standing in this third definition is a global one.

CBA step 3 has to do with cataloging impacts. The JK analysis includes (1) the additional resources (embodied in the special windows) builders use in complying with the code, (2) the reduction in energy used by households with better windows, and (3) the reduction in pollution associated with producing less energy. Pollution from energy production negatively affects people in the vicinity (such as people who suffer from breathing sulfur dioxide produced during electricity generation) and also people far away from where the electricity generation took place (such as smaller catches for fishermen because of ocean acidification caused by climate change.) Other potential impacts that Jacobsen and Kotchen did not catalog include the impact of more or less comfortable indoor temperatures; this is an impact that was included in Fowlie et al. (2018).

CBA step 4 has to do with predicting impacts. Empirical training in causal inference is critical to predicting impacts. The preceding chapters of this book have all emphasized the fact that correlation is not causation. In the crucial step 4, we need to know what impact was actually caused by the policy. It is not enough to know that energy use was lower in homes built after energy codes were strengthened, because it is possible other things changed along with regulations. For example if for some reason homes were smaller on average after the codes were strengthened, it might appear that the codes were responsible for an observed lower energy use.

One technique analysts use to estimate impacts is to get data and estimate impacts themselves. Preceding chapters described one way to do this, by using the ACS data, along with techniques like regression control, difference-in-differences, and instrumental variables. But this way is no small task. An easier way to estimate impacts, which still requires the ability to sort out correlation from causation, involves using *the literature*. By the literature, I mean all of the studies that have previously been published on a particular topic. An analyst who searches these studies will find estimates of impacts that have been produced by others. Because step 4, impact estimation, is such a crucial step in any CBA, I discuss methods for it in more detail in the text box below.

Impact Estimation: Do-It-Yourself, or Plug-in Values?

Analysts can estimate impacts using data themselves, or they can use estimates from the literature in a so-called plug-in approach. The ideal way an analyst would estimate impacts themselves is to conduct a randomized experiment. For example, Fowlie et al. (2018) use data from a wide-scale weatherization experiment, which randomly determined which homes to weatherize and later compared energy use across all homes in the experiment. As discussed in Chapter 1, randomized experiments are sometimes held as the gold standard technique for estimating causal effects, because they enable a researcher to assign treatment in a way that is uncorrelated with characteristics of participants.

Most of the time, however, experiments are infeasible because of their cost. Therefore, economists often have to rely on observational data. Jacobsen and Kotchen (2013) use utility billing data, as well as data on characteristics of the homes (such as square footage and number of bathrooms). They find that homes built just after the date that energy codes were strengthened use less energy compared to observationally identical homes built just before. Is this a compelling way to estimate the causal effect that building energy codes have on energy demand?

Arik Levinson (2016) argues not necessarily. Newer homes use less energy for reasons apart from their design, and Levinson (2016) argues Jacobsen and Kotchen conflate home vintage with home age. In a follow-up to the JK analysis, Matthew Kotchen (2017) finds evidence suggesting Levinson (2016) was correct, with regard to electricity at least, as he found energy codes were not responsible for reducing electricity demand. However, Kotchen (2017) does find that the savings from natural gas persisted and were twice as large as he and Jacobsensss had found in their 2013 analysis.

The econometric literature that estimates the impact of energy codes on energy demand is rich and evolving and reviewing it all is beyond the scope of this chapter. Table 7.2 lists nine recent studies that are all at least somewhat comparable. Care must be taken in comparing the results summarized in this table, however, because the studies use different approaches and cover different study areas. Sometimes, a single study will provide the best estimate of an impact to use in a CBA. In other situations, averaging impacts may be appropriate. An

analyst's ability to distinguish between correlation and causation is just as important when using the plug-in method as with the do-it-yourself method.[a]

[a] One of the Questions for Review at the end of this chapter discusses a CBA of reducing elementary school class size. The debate between Krueger (2003) and Hanushek (2003) highlights econometric considerations with the plug-in method, as well as the importance of Step 4, Impact Estimation, more generally.

Jacobsen and Kotchen used residential billing data to estimate regression control models that found that the change in Florida's energy code caused

Table 7.2 Estimating impacts through literature review

Study	*Finding*	*Area*
Aroonruengsawat et al. (2012)	Energy codes reduced electricity consumption by 0.3–5.0% depending on the state	USA
Koirala et al. (2013)	Energy codes reduced electricity expenditures by 1.8% and natural gas expenditures by 1.3% on average	USA
Jacobsen and Kotchen (2013)	A revision to Florida's energy codes in 2002 lowered electricity consumption by 4.3% and natural gas consumption by 6.7%	FL
Kotchen (2017)	A revision to Florida's energy codes in 2002 lowered electricity consumption by 0.0% and natural gas consumption by 13.5%	FL
Costa and Kahn (2011)	Homes built in California in the 1980s do not use significantly less electricity than homes built in the 1970s, ceteris paribus	CA
Holian (2020)	Homes built in California in the 1980s use around 2% less electricity than homes built in the 1970s, ceteris paribus	CA
Novan et al. (2017)	Homes built in Sacramento just after California adopted energy codes in 1978 use 1.6–2.6% less electricity, than those built just before	CA
Levinson (2016)	Homes built in California just after the adoption of energy codes use 0% less electricity and 5% less natural gas than homes built before, but difference is insignificant	CA
Fowlie et al. (2018)	The Weatherization Assistance Program reduced energy consumption by 10–20% but the costs of weatherization substantially outweighed the present discounted value of the private and social benefits	MI

Notes on table These nine studies use various methods to estimate the impact of energy codes on energy use

electricity consumption to fall by 48 kilowatt-hours (kWh) per month, and natural gas consumption to fall by 1.5 therms. They then use the literature to find so-called plug-in values to estimate the size of reduced emissions. The four categories of emissions they include are carbon dioxide (CO2), sulfur dioxide, nitrous oxide, and particulates. Emissions factors are numbers drawn from the literature that are used to estimate the emissions in each category produced in response to energy use. For example, JK cite a study that found burning 1 therm of natural gas generates 0.006 tons of CO2. If households reduce natural gas use by 1.5 therms per month, CO2 emissions will fall by 1.5×0.006, or 0.009 tons of CO2 monthly.[5]

CBA step 5 involves monetization—assigning a dollar amount to the impacts to represent its social value. Recall the JK analysis assumes three impacts. First, the stricter energy codes require builders to use low-E windows, and monetization involves valuing the additional resources that go into producing these windows. The JK analysis finds an estimate indicating the low-E windows are 10% more expensive than non-low-E windows, and calculate that the change to the code has added between **$675** and $1,012 to overall construction costs for a standard home. (Numbers highlighted in bold are used in the equation below.) Note that the increase in construction costs might not exactly correspond to the social costs of the resources. For example, imagine that the window manufacturing company is a monopoly and there is only a trivial increase in their cost of producing low-E windows. In that case the higher price paid by builders to the window manufacturer would be a transfer from the builder to the window manufacturer, not a social cost. Now, it is unlikely that the window producer is a monopoly, and the technique JK adopt for monetizing the impact is reasonable, but I construct this example to illustrate that there are cases where a researcher cannot just use market prices in the monetization step.

The second impact is reduced energy use. To monetize it, JK multiply the energy savings (48 kWh of electricity and 1.5 therms of natural gas per month) by the marginal price an average household pays (14.6 cents per kWh for electricity, and $1.22 per therm for natural gas) to arrive at annual energy savings of **$106**.[6] As with the construction impact discussed in

[5] Emissions factors for the other pollutants can be found in `script14.R`.

[6] The details behind this calculation are 48 kWh $\times$ 12 months $\times$ 14.6 cents = $84 annual electricity cost savings. For natural gas, the calculation is 1.5 therms $\times$ 12 months $\times$ $1.22, = $21.96. Adding these together, $22 + $84 = $106, which is the value of the annual energy savings.

the preceding paragraph, this seems to be a reasonable way of monetizing the social value of the saved resources, but it may not be perfect. For example, if the energy price consumers pay incorporates taxes, then the price consumers pay will overstate the social value of the resource savings, because part of the price is a transfer rather than a resource cost.[7]

The third set of impacts to be monetized are the four types of emissions. CO2 causes climate change and the other three are associated with public health problems (particulates—essentially soot—can cause asthma, for example). Earlier I discussed how Jacobsen and Kotchen estimate that CO2 emissions fall by 0.009 tons each month, or 0.108 tons annually, because of reductions in natural gas use. The social cost of carbon has been calculated by Nordhaus (2017) as $31 per ton of CO2 (in 2010 dollars), thus one way of monetizing the reduction in natural gas use is by multiplying 0.108 by $31, yielding an annual climate change mitigation benefit of $3.35. Jacobsen and Kotchen do not report the marginal damage figures they use for carbon, but it is possible to calculate these values from the information they do present. It turns out that they used a low estimate of $7.68 and high estimate of $93.70, in 2009 dollars; thus the $31 figure from Nordhaus lies on the lower end of the range they considered.[8] Hence the high-end estimate of the social value of reduced carbon emissions from natural gas is 0.108 tons times $93.70, or $10.12 annually. The JK analysis applies different marginal damage estimates to each pollutant and each fuel source, and finds that all together, reductions in the four types of emissions, owing to a household's lower electricity and natural gas demand, are valued at between $14.15 and **$84.84** annually.

CBA steps 6 and 7 can be combined. Discounting accounts for the fact that a dollar saved next year is not as valuable as a dollar saved now. *Net Present Value* (NPV) is the most widely used of several decision criteria seen in CBA. In fact, because the JK analysis was strictly speaking not a CBA,

[7] It is a subtle point, but the reason monetizing with market prices is appealing is because neoclassical perfect competition theory teaches us that price equals marginal cost. However perfect competition is a theoretical condition not always met in the real world, and there are times when the use of market prices is not appropriate. In these cases, analysts have to be creative in calculating a so-called *shadow price*, which is simply the true social value of an impact.

[8] Their Table 5 indicated the low- and high-end figure for carbon dioxide reduction due to fall in natural gas use was $0.83 and $10.12, respectively. Thus the marginal damage figures are backed out as follows: $0.83/0.108 = $7.68, and $10.12/0.108 = $93.7. Subsequently, I verified with the authors this was the marginal damage figure they used.

Jacoben and Kotchen do not present NPV calculations. Instead they discuss three different types of payback periods. Besides NPV and the payback period, other decision criteria one sometimes encounters include the internal rate of return, and the benefit/cost ratio. However, there are several advantages to NPV that make it the most widely used and accepted decision criterion (Fuguitt and Wilcox 1999). If NPV is positive, this indicates that the investment, policy, project, or program produces more benefits than costs over the life of the policy.

Using all the numbers highlighted in bold above, it is possible to calculate the NPV of the change in Florida's energy codes, for a representative household in Gainesville, FL as:

$$NPV = -675 + \sum_{t}^{T} \frac{106 + 84}{(1 + r)^t}.$$

We can write this another way by specifying a *time horizon*. If $t = 1$ and $T = 10$ it means the time horizon is ten years. Then we can express NPV using the equation below, which is less compact but avoids the use of the summation operator $\sum$,

$$NPV = -675 + \frac{106 + 84}{1 + r} + \frac{106 + 84}{(1 + r)^2} + ... + \frac{106 + 84}{(1 + r)^{10}}.$$

In both equations, \$675 is the low-end estimate of the social cost of the low-E windows which is an upfront payment incurred today, \$106 is the estimate of the annual social benefit of energy resource savings, and \$84 is the high-end estimate of the annual social benefit of the avoided emissions. Because this calculation uses both the low-end cost estimate and high-end benefit estimates, it can be said to be a *best-case scenario* NPV. There are three parameters in this equation, two related to the time horizon (t and T) and the discount rate r, which effects how valuable future benefits are in the present. Like other decisions in CBA, the choice of a discount rate can be highly philosophical, but in practice analysts usually adopt a market interest rate.

The analyst selects the time horizon by choosing t and T. We could base the end of the time horizon T on the effective life of the low-E windows. Windows are long-lived durables, and arguably T should be substantially higher than 10, perhaps even as high as 50 or more. The JK analysis cites Stansel et al. (2007) who report the average ownership tenure in Florida as

11.5 years. I selected a time horizon of 10 above because it is close to this figure of 11.5 years, and because as a whole number it is convenient for purposes of illustration. A longer time horizon will lead to a higher NPV here, and I consider the effect of selecting different end periods below as part of the sensitivity analysis. Regarding the beginning of the time horizon, benefits will be realized once the house is built and occupied, but by starting with $t = 1$, the NPV calculation assumes that benefits are realized at the end of every year; we assume benefits are realized at the beginning of each year by setting $t = 0$.

Assuming a discount rate of 5% (so $r = 0.05$), the best-case NPV estimate is \$792.[9] The fact that NPV is positive indicates that energy codes that require low-E windows are a good social investment.[10]

CBA step 8, sensitivity analysis, refers to determining how the NPV estimate changes when one of the assumptions or estimates that went into the equation is changed. By calculating payback periods for three different definitions of standing, Jacobsen and Kotchen do present some sensitivity

[9] NPV can be computed using R, in a spreadsheet, or with and pencil and paper using the following handy formula: $-675 + 190 * [(1/.05) - (1/.05)/(1.05^{10})]$. There is an intuition behind this term in brackets. The term $1/r$ is the present value of a dollar received every year forever (an annuity that pays out forever is called a perpetuity), which is \$20 when $r = 0.05$. The second term in brackets is the present value of \$20 ten years from now, which is 12.28. So the term in brackets can be thought of as the present value of a \$1 perpetuity that is taken away in ten years. This is \$20 minus \$12.28 or \$7.72. Given there are \$190 in benefits every year for ten years, we multiply \$190 by 7.72 to find \$1,467, the present value of benefits. From this we subtract \$675 which is the initial upfront costs to find a present value of 1,467 minus 675 or \$792.

[10] How does this NPV estimate compare with the decision criteria presented in the JK analysis? Jacobsen and Kotchen presented three criteria, the first of which is a private payback period, which is calculated as the upfront costs of \$675 divided by the annual savings of \$106 which comes to 6.37 years. This is the amount of time it would take a homeowner to recover their investment in the thicker windows. This criterion assumes a zero discount rate and does not account for impacts on third-parties, but is easy to interpret. The second criterion could be called a global social payback period, which is the upfront costs of \$675 divided by \$190, the sum of private and social benefits, and comes out to be 3.5 years. Third, Jacobsen and Kotchen recognize that, "...one might argue that the benefits associated with a lower CO2 emissions should not be considered...as they are likely to occur for the most part outside the policy jurisdiction" (p. 47). Excluding CO2 reduction benefits reduce the value of emissions reductions from \$84 to \$22, and what could be called a national social payback period rises to 5.3 years (\$675 divided by \$128, where \$128 is the sum of \$106 and \$22.) A decision maker (homeowner, policy maker) would have to somehow determine a cutoff value for the payback period to make a decision regarding low-E windows.

analysis. They do not discuss how sensitive their findings are to changes in other assumptions.

In this and the next two paragraphs, I present some examples of further sensitivity analysis. The payback periods considered in the JK analysis were based on best-case assumptions, and so I first recalculate my NPV figure using the *worst-case* figures. Recall that the calculations above used the low-end cost estimate of the low-E windows of $675, but the high-end estimate was $1,012.[11] In addition, we used the high-end estimate of the value of emissions reductions of $84, but the low-end estimate was $14. A worst-case NPV calculation would simply replace $675 with $1,012 and $84 with $14 in the equations above. With a discount rate of 5%, the worst-case NPV estimate is −$85. This negative value indicates that the discounted value of social benefits is not enough to justify the upfront costs of low-E windows.

Another assumption is the impact of the energy code changes on energy demand. In the empirical part of their study, Jacobsen and Kotchen find it to be 48 kWh per month for electricity, and 1.5 therms for natural gas. However in follow-up work using more recent data from the same study area, Kotchen (2017) finds that there are no electricity savings, but natural gas savings are about double. In terms of the NPV calculation above, natural gas savings was $22 and electricity savings $84, for a combined energy savings of $106. If we double natural gas savings and ignore electricity savings, energy savings under the revised impact estimates are only $44. In addition, social benefits of avoided emissions under these revised impact estimates range from $1.84 to $20.74. Recalculating NPV under these assumptions, I find best- and worst-case NPV estimates of −$175 to −$658, respectively. Both best and worst-case NPV figures are negative under Kotchen's (2017) revised impact estimates.

Of course, the −$175 to −$658 NPV figures presented above use the 10-year time horizon, which—as mentioned above—might be too short. As a final check on the sensitivity of these estimates, I note that with a 50-year

[11] Florida's energy code gives the builder flexibility about how to meet the energy use requirements specified in the home. If there is a design change that enables the builder to comply with the code more cheaply than by using low-E windows, the builder could select that design feature instead. This means that the $675 figure might overstate the actual cost of compliance—though I still refer to $675 as the low-end estimate.

time horizon and using the revised impact estimates from Kotchen (2017), the best and worst-case NPV figures are $507 and −$175, respectively.[12]

What can we conclude from examining the effect of alternate sets of assumptions on the NPV estimate? The NPV estimates are quite sensitive to the assumptions. CBA does not give us a clear answer in this case. While it may seem as if CBA provides a non-answer, the results do suggest that Florida's changes to its energy codes were not obviously good or bad. Then again, the sensitivity analysis does draw our attention to the fact that the marginal damage figure we use for carbon dioxide reductions is a key driver of whether the NPV is positive or negative. Assumptions about how carbon reductions impact climate change to a large extent determine whether the policy is efficient or not.

CBA step 9 entails making a recommendation. Jacobsen and Kotchen never make an explicit policy recommendation in their analysis, but implicit in it might be a suggestion that Florida policy makers were correct to strengthen the energy codes in 2001. The authors never actually state this, but it is not hard to imagine a reader interpreting their results as encouragement to further strengthen energy codes in Florida, or to replicate Florida's changes in other states in similar climate zones. However, as we have just seen, the revised empirical estimates of the policy's impact show that the case for energy codes is weaker than Jacobsen and Kotchen initially found.

My reexamination of the JK analysis suggests that Florida's stricter building codes do not clearly pass a cost–benefit test. Of course, there is always room for strengthening any analysis. Strictly speaking Jacobsen and Kotchen set out to calculate social payback periods for a representative household, not to carry out a social CBA. We have seen that it is possible to recast their analysis as a simple CBA just by specifying a time horizon and calculating NPV with the figures they provide. Thus on one hand, the analysis they carry out is very close to a CBA. On the other hand, had their goal been a comprehensive CBA they likely would have (among other things) factored in other impacts, such as the administrative costs of creating and enforcing energy codes. Recent work by Novan et al. (2017)

[12] One further consideration is worth highlighting. The analysis described above was for a representative home. If all homes in the study area are basically the same, we could simply multiply the NPV for a single home, which is what we calculated above, by the number of homes. A more careful analysis would have to account for the fact that homes differ. It turns out, many homes in Florida do not use natural gas at all; evidence from the ACS suggests only about 25% of recently constructed homes use natural gas.

adopts a different approach to valuing the cost of complying with energy codes, and in their CBA of California's energy codes these authors find evidence suggesting that the initial codes likely do pass a cost–benefit test (that is, NPV is likely positive). The question of the efficiency of building energy codes remains an active area of scholarship.

Conclusion

This chapter introduced CBA as a tool for guiding public policy decision-making. It showed how CBA incorporates empirical estimates from research studies like those described in earlier chapters, and how it incorporates environmental and public health impacts.

It then presented a case study of a CBA of a change made to Florida's building energy codes in 2002. In reconsidering the Jacobsen and Kotchen (2013) analysis as a CBA, I calculated NPV, which is the most conventional decision criterion in CBA, under best- and worst-case scenarios, and I also updated the analysis to account for new policy impacts estimated in Kotchen (2017). I find that while NPV is positive in the best-case scenario, it is negative in the worst-case scenario. When the updated impact estimates are used, both best- and worst-case NPV figures are negative. With a longer time horizon and updated impact estimates, the best-case assumptions result in positive NPV while the worst-case assumptions result in negative NPV.

This case study shows how CBA can be applied in the specific setting of home energy codes. In addition, because all CBAs follow the same steps, the case study also illustrates what CBA is generally, so it can be applied to any of the areas discussed in other chapters of this book. Some of the Questions for Review for this chapter ask a reader to consider how CBA could be applied in other areas, including education and immigration.

This chapter also emphasized that an analyst with empirical training in the causal inference techniques that are the main focus of this book will do a better job at the crucial step of impact estimation. To do CBA well, and to understand what it is, and maybe more importantly what it is not, requires an analyst to have a mix of skills (including empirical skills), a grasp of neoclassical economic theory, and a familiarity with financial calculations such as NPV and inflation adjustments. It also requires a healthy dose of critical thinking skills, both in terms of cataloging impacts, and selecting studies for the literature review that contain the most appropriate estimates to plug in at various points in the analysis.

I have stressed that CBA is not a panacea for solving all social problems. As a practical matter, sometimes our estimates of policy impacts in CBA could be very far off. This is sometimes called the knowledge problem. Philosophically, CBA can make recommendations that are at odds with distributive justice or rights. CBA relies on concepts of willingness to pay and value of a statistical life to measure benefits and costs, and at times these may not accurately reflect human welfare. Despite its limitations, the great virtues of CBA are to force comprehensive, rational decision-making, that accounts for the preferences of all members of society.

Keep in mind, the perfect CBA, just like the perfect empirical study, has yet to be written. There is always room for improvement. Ultimately decision makers have multiple criteria beyond NPV to consider, but the consequentialist underpinning of CBA deserves a place at the table in any major public policy decision or debate.

Key Terms

Cost–Benefit Analysis (CBA)	The value of a statistical life	Economic analysis
Economic Impact Analysis (EIA)	Fiscal Impact Analysis (FIA)	Standing
The literature	Shadow price	Net Present Value (NPV)
Time horizon	Best-case scenario	Worst-case scenario

Questions for Review

1. Read the second half (roughly pp. 49–60) of Krueger (2003). In estimating the impact of smaller class sizes on test scores, does Krueger use a weighted average of many estimates from the literature, or does he select one value to plug in? Why did he make this decision? In monetizing the benefit of increased learning on future wages does Krueger use a weighted average of many estimates from the literature, or does he select one value to plug in? Krueger reports the internal rate of return (IRR) and, like Jacobsen and Kotchen, does not calculate NPV. Following the list of nine steps to a CBA from Boardman et al. (2017), describe the main features of the Krueger

CBA. Replicate the calculations, and then report the best-case and worst-case estimates of the NPV.

2. Read Fowlie et al. (2018). Describe any similarities and differences between the articles by Jacobsen and Kotchen, Krueger, and Fowlie and Greenstone, in terms of Step 5, impact estimation. In particular, note whether the authors use observational or experimental data.
3. Consider immigration. Read Blau and Mackie (2017). Compare and contrast fiscal impacts ("Immigrants will be a drain on the welfare state" or "unauthorized immigrants contribute to sales and other tax revenue") with economic impacts ("immigrants take jobs" or "immigrants further the division of labor"). To decide what we should do, we need a well-defined objective. What are the objectives in Fiscal Impact Analysis and Economic Impact Analysis?
4. Consider again immigration. Read the study by Allen et al. (2019), titled, "Border Walls." Is this a CBA, EIA, FIA, or some other kind of economic analysis?

References

Allen, Treb, Cauê de Castro Dobbin, and Melanie Morten. "Border walls." No. w25267. National Bureau of Economic Research, February 2019.

Aroonruengsawat, Anin, Maximilian Auffhammer, and Alan H. Sanstad. "The impact of state level building codes on residential electricity consumption." Energy Journal-Cleveland 33, no. 1 (2012): 31.

Blau, Franeine D., and Christopher Mackie, eds. The economic and fiscal consequences of immigration. National Academies of Sciences, Engineering, and Medicine. Washington, DC: The National Academies Press, 2017. https://doi.org/10.17226/23550.

Boardman, Anthony E., David H. Greenberg, Aidan R. Vining, and David L. Weimer. Cost-benefit analysis: Concepts and practice. Cambridge University Press, 2017.

Chaudhuri, Anoshua, and Susan G. Zieff. "Do open streets initiatives impact local businesses? The case of Sunday Streets in San Francisco, California." Journal of Transport & Health 2, no. 4 (2015): 529–539.

Costa, Dora L., and Matthew E. Kahn. "Electricity consumption and durable housing: Understanding cohort effects." American Economic Review: Papers & Proceedings 101, no. 3 (2011): 88–92.

Culhane, Dennis P., Stephen Metraux, and Trevor Hadley. "Public service reductions associated with placement of homeless persons with severe mental illness in supportive housing." Housing Policy Debate 13, no. 1 (2002): 107–163.

Fowlie, Meredith, Michael Greenstone, and Catherine Wolfram. "Do energy efficiency investments deliver? Evidence from the weatherization assistance program." The Quarterly Journal of Economics 133, no. 3 (2018): 1597–1644.

Fuguitt, Diana, and Shanton J. Wilcox. Cost-benefit analysis for public sector decision makers. Greenwood Publishing Group, 1999.

Gillingham K. Rebound Effects. In: Palgrave Macmillan (eds.), The New Palgrave Dictionary of Economics. London: Palgrave Macmillan, 2014. https://doi.org/10.1057/978-1-349-95121-5_2875-1.

Hanushek, Eric A. "The failure of input-based schooling policies." The Economic Journal 113, no. 485 (2003): F64–F98.

Holian, Matthew J. "The impact of building energy codes on household electricity expenditures." Economics Letters 186 (2020): 108841.

Jacobsen, Grant D., and Matthew J. Kotchen. "Are building codes effective at saving energy? Evidence from residential billing data in Florida." Review of Economics and Statistics 95, no. 1 (2013): 34–49.

Koirala, Bishwa S., Alok K. Bohara, and Hui Li. "Effects of energy-efficiency building codes in the energy savings and emissions of carbon dioxide." Environmental Economics and Policy Studies 15, no. 3 (2013): 271–290.

Kotchen, Matthew J. "Longer-run evidence on whether building energy codes reduce residential energy consumption." Journal of the Association of Environmental and Resource Economists 4, no. 1 (2017): 135–153.

Krueger, Alan B. "Economic considerations and class size." The Economic Journal 113, no. 485 (2003): F34–F63.

Levinson, Arik. "How much energy do building energy codes save? Evidence from California houses." American Economic Review 106, no. 10 (2016): 2867–2894.

Manning, Matthew, Shane D. Johnson, Nick Tilley, Gabriel T . W. Wong, and Margarita Vorsina. Economic analysis and efficiency in policing, criminal justice and crime reduction: What works? London: Palgrave Macmillan, 2016.

Nordhaus, William D. "Revisiting the social cost of carbon." PNAS 114, no. 7 (February 14, 2017): 1518–1523.

Novan, Kevin, Aaron Smith, and Tianxia Zhou. "Residential building codes do save energy: Evidence from hourly smart-meter data." UC Davis (2017).

Stansel, Dean, Gary Jackson, and Howard Finch. "Housing tenure and mobility with an acquisition-based property tax: The case of Florida." Journal of Housing Research 16, no. 2 (2007): 117–129.

Learning Goals for Appendix A

1. Download ACS data from IPUMS-USA.
2. Install the R and R Studio statistical software programs; list and install required packages.
3. Analyze ACS microdata by running one of the R scripts on this book's companion website.
4. List and describe best practices in the analysis of ACS data, and in replicating studies that use it.

M. J. Holian, *Data and the American Dream*,
https://doi.org/10.1007/978-3-030-64262-4

Appendix A: Open Access to Data, Software, and Code

To this day I have a greenish-gray speck of pencil graphite embedded in my right hand. Once as a child in the 1980s I was having trouble with my math homework and, in order to demonstrate my frustration, I stabbed the homework sheet with my pencil, accidentally plunging it through the paper and into the soft tissue of my right hand. The showing was more dramatic than I intended, and I ended up in the emergency room with a doctor scraping the graphite from my wound. Today, that spot on my hand serves as a reminder that I can overcome challenging math and technical problems, even if they are not easy for me.

This appendix is intended to guide someone, who may have only minimal prior experience with data analysis, in using statistical software programs. The three steps to do this involve: (1) downloading the master data file that underlies nearly all of the statistics discussed in this book, (2) installing the free R and R Studio software I used to analyze the data, and (3) using the "R scripts," which are code or "analysis files" (available on this book's companion webpage) that run with the R software, and that produce the statistics discussed in the chapters. Once a reader has accomplished these three steps, a whole world of possibilities will be at their fingertips because, as highlighted throughout this book, it is often easy to modify just one line of code (e.g., the line containing occupation code in the lawyer earnings example from Chapter 1) to determine some previously unknown statistic (such as most popular college majors or average earnings by major for software developers, as we saw at the start of Chapter 4).

M. J. Holian, *Data and the American Dream*,
https://doi.org/10.1007/978-3-030-64262-4

This appendix also contains something for professionals. In the final section, I describe some lessons I learned in writing this book, from obtaining replication files to empirical best practices.

Although only minimal prior experience is required to run and modify the scripts associated with this book, learning to use data analysis tools requires careful attention to detail, and a healthy dose of patience. Often, we learn best by doing. Thus to learn R programming, I suggest finding a project that requires you to do original data analysis, such as a term paper or research paper. My approach to teaching programming, or "coding" skills, is thus project-based. I also think it's important to have traditional textbooks, guidebooks, and course work, and I discuss these resources below.

In my own research, typical coding tasks include: producing subsamples of data from a larger sample of raw data (we might have data on all types of workers, but only calculate average earnings for the subset of practicing lawyers), recoding variables in the raw data so they can be used in the analysis (such as transforming a categorical variable into one or more binary variables), and then using the "cleaned" variables and refined subsamples to estimate statistics (such as mean lawyer earnings by college major, medians, or regression coefficients).

All of the programming I have done for this book involves common tasks. I did not know R programming when I began this project. I used a commercial (not an open-source) program called Stata, which is widely used among economists. My training in economics was probably typical in that it did not include formal training in programming, so most of what I know about programming I have taught myself.

Today my pedagogical approach relates to my personal experience in having had to "go it alone" with programming. Most of my coding questions have been asked and answered somewhere online before, and this gives me confidence that I will be able to complete whatever new task has popped up. Going back to patience, I know it can take time to solve a new coding challenge, and that deciphering error messages, reading documentation files, and careful Internet searching are necessary strategies. I do coding in the service of research, and in my view coding is problem-solving. More than having vast coding knowledge, successful programming often requires creativity and resourcefulness.

A reader who is able to run my R scripts (I also provide links to "do files," which are what analysis files are called by Stata users) following the steps detailed in this appendix will actually learn more than just programming.

They will also be exposed to research methods more broadly. Of course there's more to research than replication and extension, such as learning institutional details and selecting the research question in the first place. But by downloading the full sample of raw data from the original source, and seeing line by line how the estimates were produced, a reader will be well-positioned to think of ideas for original research. Ideas for building off of the studies discussed in this book were provided in review questions at the end of each of the preceding chapters.

This book is a good launching off point for someone who wants to learn programming or do econometric research, but someone whose work involves programming will consult thousands of articles, books, online discussion boards, and even Facebook and Twitter posts on some aspect of programming throughout their careers. Which resources should you consult for help? You must use judgment here, but I will offer a few suggestions.

The most important resource you should have is probably a comprehensive econometrics textbook. I wrote this book in part to be used as a supplemental text in introductory undergraduate econometrics courses, or in a graduate course that emphasizes research methods and writings. You don't find a lot of discussion about how econometric research is actually done in econometrics textbooks, but they do cover the details of hypothesis testing and other important topics that I have discussed at an intuitive level. Two textbooks that I use and recommend are *Real Econometrics* by Michael Bailey and *Introductory Econometrics* by James Stock and Mark Watson.

Not only do these books cover concepts that I have not emphasized, they are also both associated with free and valuable resources, such as data and analysis files hosted on their companion webpages. One can also find resources produced by the community of educators who use them. The *Guide to R: Data Analysis for Economics* by Bill Sundstrom and Michael Kevane has been a great help to me, and is associated with a library of R scripts that reproduce the estimates shown in Stock and Watson's textbook. It is available to download for free.[1] The free online book *R Companion to Real Econometrics* by Tony Carilli reproduces the analysis in Bailey's

[1] https://rpubs.com/wsundstrom/home. Another free resource that reproduces the analysis in Stock and Watson (2011) using R is *Introduction to Econometrics with R* by Christoph Hanck, Martin Arnold, Alexander Gerber, and Martin Schmelzer. https://www.econometrics-with-r.org/.

textbook. I too have contributed teaching resources to the community of educators using the Bailey textbook.[2]

I have also used in my courses a book by Josh Angrist and Steve Pischke titled *Mastering 'Metrics: The Path from Cause to Effect*. The first chapter of this book is available for free on the publisher's webpage, and contains an appendix on statistical inference and difference in means testing in experimental and non-experimental data that I highly recommend. This book doesn't contain a glossary or other traditional textbook elements, but it does have a website with Stata replication files. Josh Angrist appears in some excellent videos produced by Marginal Revolution University, and on my blog I've shared my syllabus and some of the resources I've developed for using their book in the classroom, including R scripts.[3]

In the Economics Department at San Jose State University where I teach, we recently created a two-course sequence in econometrics in our undergraduate program, designed around Angrist and Pischke's *Mastering Metrics* and Bailey's *Real Econometrics*. The idea behind the sequence is that after being exposed to basic econometrics and research methods in the first course, a student can complete an original econometric research project in the second. The replicate and extend approach, discussed in Chapter 2 and throughout this book, is a great fit for courses like these. I have also had graduate students in a research-focused, introductory econometrics course, replicate studies that use the ACS. This course used Stock and Watson's *Introductory Econometrics* as the main textbook. I also used sections from more advanced books like Angrist and Pischke's *Mostly Harmless Econometrics* and Scott Cunningham's *Causal Inference: The Mixtape*, but the journal articles students were replicating served to provide more of the advanced content for the course. The use of replication exercises in teaching is becoming increasingly common across the social sciences.[4]

My contention is that you can best learn programming by doing, and you will develop knowledge and intuition about things like "data structures" that are covered in traditional computer science courses along the

[2] Files including two R scripts can be downloaded from: http://mattholian.blogspot.com/2020/10/analyzing-donut-consumption-in.html.

[3] My teaching resources for *Mastering Metrics* can be found at http://mattholian.blogspot.com/2015/01/econometrics-and-kung-fu.html. The link to the videos is https://mru.org/mastering-econometrics.

[4] For more on this see: Höffler (2013) in economics, Janz (2016) in political science, and Jekel et al. (2020) in psychology.

way. That said, traditional training in data science, computer science, and econometrics would allow you to avoid learning things on your own all the time, and so I therefore do recommended formal coursework. The difficulty with some of this training is that it can be hard to get motivated to learn things you "might use sometime."

The more you code the better equipped you become to solve future coding challenges. Adopting a growth mindset is important. Someone with a fixed mindset believes you are either born to code or you are not; someone with a growth mindset says things like, "I am not good at coding...yet." Before turning to the technical details of this appendix in the following four sections, I'll end this introductory section by suggesting one more open-source software program a reader may wish to learn, and provide a final piece of motivation.

The software I used to write this book is LaTeX, sometimes written LaTeX. I pronounce it "Lay Tech" but there seem to be differing opinions on this. This software is frequently used in technical reports written by mathematicians, engineers, and economists. Like R for statistical analysis, and QGIS for maps (I discussed the open-source QGIS program in footnote 2 in Chapter 1), LaTeX is free, open-source software for word processing.[5] There are always fixed costs to learning a new software language, but after my draft manuscript, which I had been writing in MS Word, crashed on me for the tenth time, I glanced down at the pencil graphite in my hand and decided it was time to learn LaTeX. The attitude I try to encourage in my students is the same attitude I strive to adopt myself.

Obtaining Data

This section describes how to obtain the raw ACS data from IPUMS (Ruggles et al. 2020). We download these data from a center at the University of Minnesota called IPUMS. Among professionals, this acronym is pronounced with a short i as in "integrated" rather than long "i" as in "i-phone." IPUMS-USA distributes U.S. Census microdata, and there are several other IPUMS divisions; as two examples, IPUMS-International distributes microdata from other countries, and IPUMS-TERRA has, "Integrated data on population and the environment from 1960 to the present." This book uses just IPUMS-USA data, but once you understand how to

[5] In fact, you can use LaTeX files right in R Studio, though I typically use a program called TexEdit to compose LaTeX documents.

use it you will be in a good position to start using other data they distribute. Someone could spend a lifetime using the IPUMS data to analyze social science and related questions.

The IPUMS website makes accessing data as easy as making an Amazon purchase. (Well, almost. There's no app yet.) First, register for a free account. Second, log on and navigate to "Get Data." Click "Add Samples" and select the years 2004–2017. This is a total of 14 samples. Next, select "Add Variables."

Adding variables will take a few minutes, as there are a few dozen variables you will have to select (42 to be exact). I would recommend you select variables by letter. The 42 variables in Table A.1 are listed in alphabetical order, so start with "A," add AGE to cart, then select "B" and add BEDROOMS, and continue like this until you have added all of the variables to your "cart."[6] The variables that appear in Table A.1 are those needed to replicate the studies listed in Table A.6, and the other statistics presented throughout the book.

After you have added the 14 Samples and 42 Variables to your cart, it's time to "check out." You do this by selecting View Cart -> Create Data Extract. Make sure to select CSV under Data Format. It tells us here that the estimated size of the file is 8939.8 MB. This is about 9 GB. You will know your data request is being processed when you see, "Your extract request 1 has been submitted." In my experience it usually takes about an hour for the IPUMS servers to process an extract, though it can take longer. After the extract is processed, the download can take an additional hour, depending on your Internet connection speed. If you log into your account and visit "My extracts" after the data file is processed, it will be available for download for a few days, after which time you'll have to resubmit your extract. IPUMS stores the information on the samples and variables we selected under "My account" indefinitely, but it does not save the large data files there for long.

After the extract is processed, you will be able to download a compressed (zipped) file in GZ file format. After you download it, unzip it (decompress it). This usually just involves double-clicking, or right-clicking and selecting the appropriate option. Snags can happen at any point, however, so if you are stuck on the decompressing stage, stay calm and find a solution because

[6] When you get to Q you will see there are no variables listed. Instead, to add QINCWAGE and QBPL data quality flag variables to your cart, use the search option to search for them by name.

Table A.1 Variables in master data file for this book

Variable	*Description*
AGE	Age
BEDROOMS	Number of bedrooms
BPL	Birthplace
BUILTYR2	Age of structure, decade
CITIZEN	Citizenship status
CLASSWKR	Class of worker
COSTELEC	Annual electricity cost
COSTGAS	Annual gas cost
COUNTYFIP	County FIPS code
CPI99	CPI-U adjustment factor to 1999 dollars
DEGFIELD	Field of degree
EDUC	Educational attainment
EMPSTAT	Employment status
FUELHEAT	Home heating fuel
HHINCOME	Total household income
HHTYPE	Household Type
HISPAN	Hispanic origin
INCBUS00	Business and farm income, in 2000 dollars
INCEARN	Total personal earned income
LABFORCE	Labor force status
MARST	Marital status
MOVEDIN	When occupant moved into residence
NCHILD	Number of own children in the household
NUMPREC	Number of person records following
OCC1990	Occupation, 1990 basis
OWNERSHP	Ownership of dwelling
PUMA	Public Use Microdata Area
QBPL	Data quality flag for Bpl, Nativity
QINCWAGE	Data quality flag for incwage, inctot, incearn
RACE	Race
RELATE	Relationship to household head
RENT	Monthly rent
ROOMS	Number of rooms
SEI	Duncan Socioeconomic Index
SEX	Sex
STATEFIP	State FIPS code
UHRSWORK	Usual hours worked per week
UNITSSTR	Type of housing structure
VALUEH	Value of owner-occupied housing
VEHICLES	Vehicles available
WKSWORK2	Weeks worked last year, intervalled
YRIMMIG	Year of immigration

few of these steps are optional. After you unzip the download, you'll see a CSV file type which should have a name like `usa_00001.csv`. In principle CSV files can be opened with MS Excel but I strongly discourage you from trying to do so with this CSV file; there are so many observations it will probably crash your computer. Hold off on opening the data file for now.

At this point, it is time to create a "directory folder." This is just a folder you create somewhere on your hard drive where we will keep the data and analysis files for your work with this book. It is also the place you can find files that R creates for you. It can be on your desktop, in your documents folder or another location. We'll refer to this folder as your "directory folder" and you can name it "Rscripts" or "Rfiles" or any name you like. We'll just move that CSV file discussed in the previous paragraph to this folder for safekeeping for now. If at this point you have a file named something like `usa_00001.csv` in a folder on your computer that you specifically created to use with R, you are mostly done with obtaining the data. It would be possible now to skip to the next section, Obtaining Software, to open and view this file. However, it is critical for a researcher to try to understand their data. The remainder of this section explains key details of the ACS data.

In addition to the CSV file, there is one other file you should download before closing the IPUMS webpage, and this is the *codebook* file.[7] On the IPUMS download screen, where you found the link to the compressed CSV file, you also saw two options for downloading the codebook. The link for "Basic" links to a CBK type file, and the "DDI" link to an XML file. If you click the "Basic" link, the codebook should open in your web browser (it works for me using Mozilla Firefox), then you can "save as" this web page as a PDF file which makes it easy to view later. If you download the codebook file from IPUMS in CBK format, using the Basic link, the CBK file format can be opened later with a text editor like Text Edit on a Mac or Notepad on a PC.

Although we selected 42 variables, we get 61 in our download. Why? There are two reasons. First variables like YEAR, CBSERIAL, PERWT, NUMPREC, HHWT are all preselected; there are ten such variables that are included whether we want them or not. Second, some variables like

[7] If you fail to save the codebook file, you can always find descriptions of the variables on the IPUMS-USA webpage; for example: https://usa.ipums.org/usa-action/variables/SEX. This page and subpages contain important information about the variables beyond codebook details.

EDUC, are distributed as two separate variables, one "basic" and one "detailed." There are nine such variables. For example, the basic variable EDUC contains 11 categories of educational attainment. This means we know roughly how much education someone completed. The variable EDUCD is the "detail" version of EDUC, included automatically whenever EDUC is selected. It contains 24 categories of educational attainment. Tables A.2 and A.3 list the nineteen variables with descriptions that are included in our extract that we did not select.

Table A.2 Ten preselected variables are automatically included with all extracts

Variable	*Description*
YEAR	Census year
SAMPLE	IPUMS sample identifier
SERIAL	Household serial number
CBSERIAL	Original Census Bureau household serial number
HHWT	Household weight
CLUSTER	Household cluster for variance estimation
STRATA	Household strata for variance estimation
GQ	Group quarters status
PERNUM	Person number in sample unit
PERWT	Person weight

Table A.3 Nine detailed version variables that are automatically included with basic versions

Variable	*Description*
BPLD	Birthplace
CLASSWKRD	Class of worker
DEGFIELDD	Field of degree
EDUCD	Educational attainment
EMPSTATD	Employment status
HISPAND	Hispanic origin
OWNERSHPD	Ownership of dwelling
RELATED	Relationship to household head
RACED	Race

Note Nine of the user-selected variables in Table A.1 have detailed versions

There are many nuances regarding these data.[8] At this point let's move to discuss the codebook in more detail, as you will need it to understand the data when we finally open the data file in the next sections.

In Chapter 1, Tables 1.1 and 1.2, I showed some examples of actual raw ACS data, and imagined what these people and households might be like. To understand the meaning of the raw data, we saw the Codebook was critical. A careful researcher will have to look through the whole codebook document to really understand what the variables measure. Here I will discuss the codebook definitions for a few variables.

The codebook for the IPUMS download described above is 5000 rows long, or 283 pages when converted into a PDF document. Thus I cannot reproduce here the full coding for each variable used in this book. Instead, in Table A.4 I provide full codebook details for some of the variables discussed in the Introduction, with partial details for two (OCC1990 and HHTYPE).

Using a codebook can be intimidating at first because of its length, but it is important to always consult the codebook, even when you think it is obvious how the variable would be coded. Take the example of the variable SEX. Does 1 represent female because female comes before male, or does 1 represent man, because man comes before woman in alphabetical order? To answer this, examine Table A.4 or consult the codebook file you downloaded from IPUMS. Next consider the example of household income (HHINCOME). If we didn't look, we might think 9999999 meant this household made just shy of ten million dollars each year, when in fact it means N/A-not applicable. There is no household income data for individuals who live in group quarters.

[8] For example not all variables are available for all years, as new questions are sometimes added to the ACS over the years. The early surveys (2001–2004) did not include city, county, or PUMA location for place of residence, place of work, or place of previous residence. College degree questions were added in 2009.

Table A.4 Codebook values for selected person and household variables

Variable	*Codebook values and interpretation*
CBSERIAL	Census Bureau household identification number
AGE	0 Less than 1 year; 1 One year; ... 135 years
CITIZEN	0 N/A; 1 Born abroad of American parents; 2 Naturalized citizen; 3 Not a citizen; 4 Not a citizen, but has received first papers; 5 Foreign-born, citizenship status not reported
RELATED	101 Head/Householder; 201 Spouse; 301 Child; 302 Adopted Child; 303 Stepchild; 401 Child-in-law; 501 Parent; 601 Parent-in-Law; 701 Sibling; 801 Sibling-in-Law; 901 Grandchild; 1001 Other Relatives; 1114 Unmarried Partner; 1115 Housemate/Roomate; 1241 Roomers/boarders/lodgers; 1242 Foster children; 1260 Other non-relatives; 1270 Group quarters member; 1301 Institutional inmates
EDUCD	1 N/A; 2 No schooling completed; 11 Nursery school, preschool; 12 Kindergarten; 14 Grade 1; 15 Grade 2; 16 Grade 3; 17 Grade 4; 22 Grade 5; 23 Grade 6; 25 Grade 7; 26 Grade 8; 30 Grade 9; 40 Grade 10; 50 Grade 11; 61 12th grade, no diploma; 63 Regular high school diploma; 64 GED or credential; 65 Some college, but less than 1 year; 71 1 or more years of college credit, no degree; 81 Associate's degree; 101 bachelor's degree; 114 master's degree; 115 Professional degree beyond a bachelor's degree; 116 Doctoral degree
SEX	1 Male, 2 Female
INCTOT	0000001 = $1 or break even, 9999999 = N/A, −$19,998 Bottom code, No Topcode.
UHRSWORK	0 N/A, 1-98 1-98 hours, 99 Top code
OCC1990	055 Electrical engineer; 103 Physical therapists; 156 Primary school teachers; 178 Lawyers; 217 Drafters; 276 Cashiers; 337 Bookkeepers, accounting clerks; 379 General office clerks; 417 Fire fighting, prevention, and inspection; 229 Computer software developers; 999 Unknown
HHINCOME	9999999 = N/A, −$19,998 Bottom code, No Topcode.
HHTYPE	0 N/A; 1 Married-couple family household; 2 Male householder, no wife present; ... 7 Female householder, not living alone; 9 HHTYPE could not be determined.
ROOMS	00 N/A, 1 one room, 2 2, ... 30 30 rooms
RENT	0000 = N/A, 0001 = No cash rent, top codes by state
VALUEH	9999999 = Missing, top code by state
VEHICLES	0 N/A, 1 1 available, 2 2, ..., 9 no vehicles available

Notes Select values only shown for HHTYPE and OCC1990
CBSERIAL values are not unique; they are reassigned every survey wave
Values indicating N/A are interpreted differently depending on the variable
As two examples: N/A for CITIZEN indicates a person was born in USA (see question wording in Fig. B.8 in Appendix B). For VALUEH a value of N/A indicates both that the person lives in group quarters, and that the home is rented

An important point is that a value of 9999999 for HHINCOME (or N/A more generally) does not indicate a nonresponse or missing value.[9] It's best to think of the examples of N/A in Table A.4 as "not applicable" rather than "not available." If the survey respondent did not answer the income question, an imputed value for HHINCOME would still appear in the data. The nonresponse would be indicated through a data quality flag variable, QHHINCOME, so a researcher could drop variables with data quality flags, a point I return to in the final section.

Obtaining Software

The CSV file we downloaded and saved to a directory folder in the last section is about 9 GB, the same size as a few high definition picture files. Although it is possible to open CSV type files with a spreadsheet program like MS Excel, trying to open a 9 GB file will probably crash your computer. If you have a decent computer (with at least 4 GB of RAM) you probably don't need a different machine, but you will need special software. This section covers downloading, installing, and using two programs: R and then RStudio. R is the program that runs the data analysis, while RStudio is an interface for R, in which we can edit scripts, see output, variables, and datasets, manage files and more. To do analysis we need to open RStudio only; it will automatically open R.

Download and install the R and RStudio software packages. These are different programs that work together, and you must install R first. The R program, sometimes called "base R" is the program that executes data analysis commands, while R Studio is an easy way to "interface" or interact

[9] A case where a survey respondent does not answer a question is referred to as item nonresponse. In these cases, the Census Bureau allocates or assigns values rather than report it as N/A. We can determine in which cases values were assigned or allocated, by selecting data quality flag variables in our IPUMS extract, in this case, the QHHINCOME variable. Two data quality variables appeared in Table A.1: QINCWAGE and QBPL. See https://www.census.gov/acs/www/methodology/sample-size-and-data-quality/ and links therein for data quality and response rate definitions and measures. Table A.4 also reveals the use of "top codes" and "bottom codes." For example, an individual that works 16 hours a day or 112 hours per week will have their hours (UHRSWORK) reported in the data as 99 because their actual value exceeds the top code. Someone who experienced a huge loss in income in the year will have their income reported in the data as −$19,998. One reason for top and bottom codes is to ensure the identity of all respondents remains anonymous.

with R. Download the R program at the Comprehensive R Archive Network(CRAN) website:

https://cran.r-project.org/

Mac users should click the Download link corresponding to the Mac operating system, and Windows users should click the download link for their operating system. (Linux users, you know what to do.) Next you need to download and install R Studio:

https://www.rstudio.com/

If your download and installation of the programs was successful, when you launch RStudio you will see it has four basic windows: Script editor (top left), Workspace (top right), R console (bottom left), and Session management (bottom right). These elements are shown in Fig. A.1.

Before you are able to run the analysis files associated with this book, you'll need to install "packages." This is unlike commercial software programs like Stata which come pre-loaded with most features you need, but is common in Open-Source software programs like R and LaTeX. Eleven packages are used across all the scripts associated with this book. You can install all packages with two lines of code, shown on lines 30–31 in the script editor (top left of Fig. A.1).[10] Delete the #, highlight the lines, and click the Run button. This will install eleven packages and may take about 15 minutes. Once you install the packages, they are on your hard drive and you will not have to install them again, so you can then make lines 30–31 comments again.[11]

Packages only need to be installed once but must be loaded (with the "library" command) every time you start R and want to use that package. Loading the package essentially means "turning it on." For example, line 39 in the script editor (top left) of Fig. A.1 contains the command to load the "car" package. The packages tab of the bottom right window shows a list of all packages that have been installed. We see there that the "car" package has been loaded, because it is checked. We can also see in Fig. A.1 that all of the code has already been run, because data frames and other value appear in the top right window, and the output (a regression results

[10] In that figure a # appears at the beginning of these lines; this is a way of "commenting out" the code to prevent installing packages twice. Beginning a line with a # prevents R from running the line, treating it as a comment. To run it we only need to delete the # first. Is not a big deal if you install packages twice, but it can be time-consuming.

[11] Some Mac users with older operating systems cannot install the "car" package, which is used extensively in the scripts discussed in the next section.

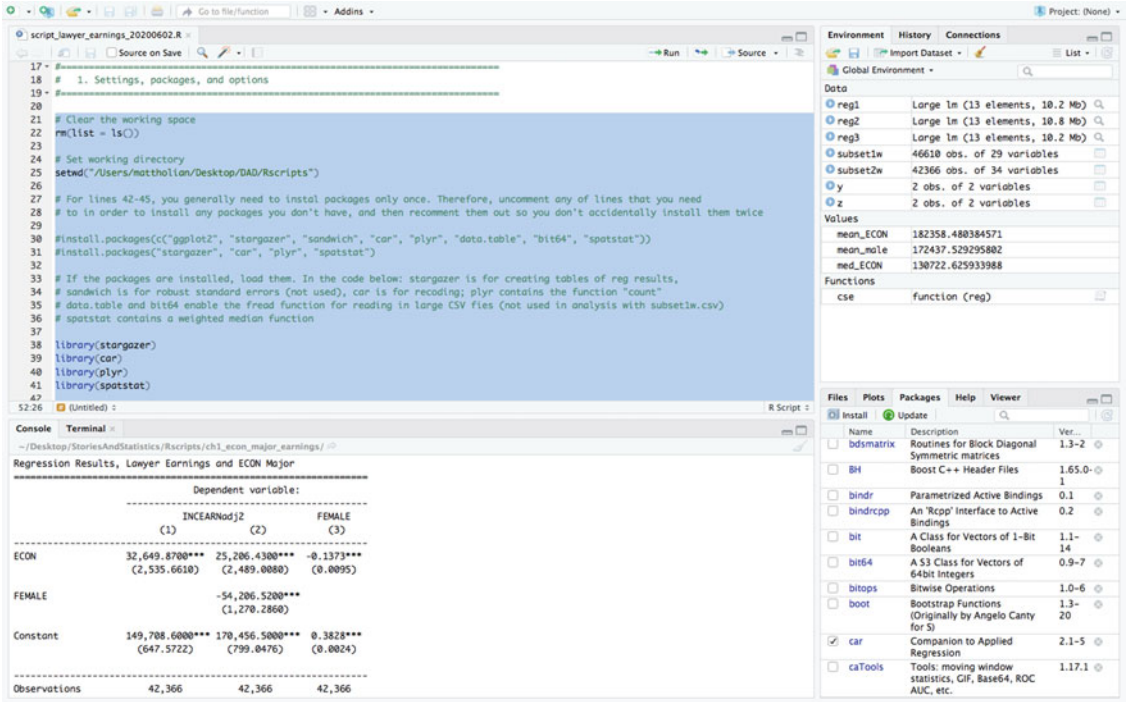


Fig. A.1 R Studio Interface showing four basic windows: Script editor, Workspace, R console, and Session management

table) appears in the console. The next section describes running the R scripts in more detail.

Using Scripts to Analyze the ACS Data in R

You now have the CSV data file in a directory folder, and you have installed the software. Below I provide a link to download R scripts–which are the analysis files–from this book's companion webpage. If you open one of these `.R` files in R Studio, they will appear in the script editor (the top left box in Fig. A.1). I have divided each of the R script files for this book into three sections: (1) Settings, (2) Data, and (3) Analysis. The list of packages needed to run the program appears in the first section, "Settings." The R scripts will load the required packages for you, but it is still up to you to make sure they are installed in advance of running the programs.

This book contains several dozen statistics, many of which are replications of those that have been previously published. Tables A.5 and A.6 describe the computer files associated with this book in two ways. First Table A.5 presents a list of all files and the chapters in which they are used. Second, Table A.6 lists replication files for the articles that served as case studies. For many cases I did not replicate the entire study, and sometimes it was just one or two key results. One of my hopes with this book is that it will help build a community of users who will finish the replications, as well as produce new replications, extensions, and original research.[12]

R scripts for this book can be downloaded from the following webpage[13]:

https://sites.google.com/site/profholian/home/dad

[12] This book's web page contains a form users can use to submit replications they have carried out of studies that use IPUMS-USA data. We plan to update this book's companion page with links to the replications submitted by the community of users of this book.

[13] I have also archived the data and analysis files for this entire book as Holian (2021) at: https://www.openicpsr.org/openicpsr/project/135661. There, you'll find the first edition of the files; any subsequent improvements or additions will initially be posted on the companion webpage, which will also host the replications submitted by the user community. OpenICPSR provides a more appropriate repository for research data for a scholarly book like this, but it will be less dynamic than the Google Sites page.

Table A.5 Code and data files to replicate statistics in this book

File	*Chapter*	*Description*
`script01.R`	1	Produces PUMA-level statistics in Table 1.3
`script02.R`	1	Winters lawyer earnings verification
`script03.R`	1	Produces bedrooms and sibling gender statistics
`script04.R`	1	Produces answers to Chapter 1 review questions
`script05.R`	2	Costa and Kahn home energy verification
`script06.R`	2	Costa and Kahn home energy reproduction
`script07.R`	2	Holian home energy verification
`script08.R`	3	Orrenius and Zavodny immigration verification
`script09.R`	4	Produces software developer earnings statistics
`script10.R`	4	Bailey and Dave health insurance verification
`script11.R`	5	Produces gig economy statistics
`script12.R`	5	Comolli and Bernardi fertiliity verification
`script13.R`	6	Holian vehicle ownership verification
`script14.R`	7	Produces Cost–Benefit Analysis calculations
`ACSmaster.RData*`	1–6	Data file described in section Obtaining Data

*All scripts, except `script05.R` and `script06.R`, use this data file. It was obtained from IPUMS and consists of the complete ACS samples from 2004–2017 for variables listed Tables A.1–A.3. Data for `script05.R` is from the *American Economic Review*, and data for `script06.R` is 2000 long-form decennial Census data obtained from IPUMS. For more details see Chapter 2, Review Questions 2 and 3, and the comments within the script files

Table A.6 Seven case studies, the samples they used, and replication files

Case Study	*Samples*	*Code File**
Winters (2016)	2009–2013 ACS	`script02.R`
Costa and Kahn (2011)	2000 Decennial Census	`script05.R`
Holian (2020b)	2013–2017 ACS	`script07.R`
Orrenius and Zavodny (2015)	2005–2006 ACS	`script08.R`
Bailey and Dave (2019)	2005–2016 ACS	`script10.R`
Comolli and Bernardi (2015)	2004, 2007 & 2010 ACS	`script12.R`
Holian (2020a)	2013–2017 ACS	`script13.R`

*All scripts except `script05.R` require the `ACSmaster.RData` data file
See the note to Table A.5 on the data required to run `script05.R`

A unique feature of this book is that each of the R scripts (except `script5.R` and `script6.R`) use the same master data file, the one downloaded from IPUMS and described in the section Obtaining Data. There we created a folder (our directory folder) and moved the unzipped CSV file we downloaded from IPUMS there. Each of the scripts contains code to transform this CSV file into an RData file, named `masterACS.RData` and

this is the "master" data file for this book.[14] The file `ACSmaster.RData` is a large file, but keep in mind it is nowhere near as large as the file would be if it included all variables and all samples available from IPUMS. It is large enough to enable us to estimate every statistic I present in this book, but it is small in comparison to the total amount of data available from IPUMS.

Download the R scripts listed in Table A.5 from this book's companion webpage, and put them in the directory folder you created. You must modify one line of each of the R scripts to indicate where your directory folder is located, because it's not the same location on your computer as it was on mine. The location of the directory folder is called its "path." To find the path, right-click on any file in the directory folder, and select "Get Info" on a Mac or "Properties" in Windows. Copy the location and paste it in the script, replacing my path with yours.[15] The precise location where the directory path appears in the script in Fig. A.1 is at line 25.

If you have successfully installed R and R Studio, downloaded and unzipped the data, moved it to a directory folder, installed packages, and modified line 25 of the R script so it refers to your working directory, the moment of truth has arrived. You are ready to run some analysis. You can run all the code in an R script all at once, but it is usually better to run the R scripts one section or sometimes one line at a time. In Fig. A.1 you see I have highlighted Section 1 of the lawyer earnings (Winters 2016 replication) script, discussed in Chapter 1, because it appears in blue (this will be gray in some print versions of this book). There is a button in the script editor labeled "Run." With the code highlighted, click Run and R will process the highlighted section.

After you highlight all code in Section 1 and press Run, what happens? You should see a lot of messages appear in the console relating to loading packages. Some messages may be in red but these are not necessarily things you need to worry about. Check the packages tab of the Session Management element of R Studio. Some packages should have check marks next to them, as in Fig. A.1, indicating they are "loaded" and ready to be used.

[14] The file `ACSmaster.RData` is available to download from this book's companion website, however, I recommend learning how to use the IPUMS download system because this book's master data file doesn't contain all the variables and samples we will ever need.

[15] Mac users can just copy and paste; PC users need to change the backward slashes to forward slashes after pasting.

Next move to Section 2. There are a number of ways to load the data. Some users have memory limitations that prevent loading the master data file, so I have made subsets available for most of the scripts. Read the comments in Section 2 of the R scripts to determine which line you need to run to load the data, depending on your situation.

Now I describe how someone can load the CSV data file they obtained from IPUMS into R Studio. The code that loads the CSV file has been "commented out" in all scripts and will not run unless you remove the # symbol. Loading this large file will take a long time, perhaps an hour. You'll see a "stop sign" symbol at the top of the console that indicates R is processing. Luckily, once you load the data in the CSV for the first time, you can save it as an RData file, and will never have to load the CSV file again. The RData file will load much faster.

After loading either the master data or a subset, a data set (or data frame) appears in the Workspace. Next, highlight and run the rest of the code in Section 2. You'll see more data frames appear, because some of the code in Section 2 creates estimation subsamples. You'll also see new variables are created. Very rarely can we use a variable in the raw data in analysis without first transforming it, in consultation with the codebook. Study the code in Section 2 to see how the raw data is recoded into variables which are used in the analysis.

Finally we come to Section 3 of the R script, Analysis. This produces results you see in the Console in Fig. A.1. These results match up with those I reported in Chapter 1. Study this code to see how R calculates statistics like means and medians, and runs bivariate and multivariate regressions, and creates a table of regression results.

Many first-time students struggle with the idea of having a directory folder, but specifying one allows you to save output automatically, and you'll know where to find it. For example in the lawyer earnings script, the frequency of lawyers with each major is produced by the line: `y=count(subset1w, 'major')`. There are a lot of different majors, so it is helpful to save the output of this command as a CSV file, which can be done with the line: `write.csv(y, file = "degreesTabulate.csv")`. After you run this code, look in your directory folder for the file `degreesTabulate.csv`. You can open it up with a spreadsheet (unlike

the data file you downloaded from IPUMS, this CSV file is small) and compare the results with those from the Winters (2016) study.[16]

Finally, what if you've followed these instructions exactly and you still cannot run one of the scripts? Most of the problems I've seen involve computers with insufficient memory. Sometimes users have enough memory but need to free up memory with the command `gc()` or allocate more memory to R using a command like `memory.limit(size=20000)`. If your machine simply does not meet the minimum requirements to run all packages with the master ACS file,[17] you have options. First, you can use the data subsets, produced from the `ACSmaster` data frame within the scripts and available on this book's companion webpage, rather than the full ACS master file. This will leave you with fewer options for extensions, but it will allow you to run scripts to verify all statistics contained in this book. Second, you can try running scripts "in the cloud" on a virtual machine. Currently, R Studio Cloud accounts are free with moderate memory restrictions. The companion webpage provides links to scripts I have set up in the cloud. While the price of R Studio Cloud could change at any time, it is likely that the prevalence of cloud computing options will continue to increase.

Finding Studies to Replicate and Extend, and Other Lessons

In the course of writing this book, I have learned several lessons about what type of study makes a good candidate to replicate. I have also picked up some best practices in analysis of ACS microdata. In the final section of this appendix, I discuss these lessons.

A study that is a good candidate for a reader of this book to replicate has three characteristics: (1) Uses microdata from IPUMS, (2) Only estimates models covered in this book, and (3) Does not use merged data (or you have replication files if it does). Many of the studies cited in the chapter meet these criteria.

[16] The stargazer package is used to create nicely formatted tables. Use the "out" argument to save a table in .html format to your directory folder, open it with a web browser, and copy and paste the image into your word processing program. The code in `script4.R`, specifically the section that carries out analysis for Chapter 1, Review Question 5, shows how to do this.

[17] I am running macOS Version 10.13.6, on a 2.2 GHz Intel Core i7 processor with 8 GB of Memory.

As scholarly journals continue to demand more from authors in terms of archived research data and code, another characteristic to consider when selecting a study to replicate is whether or not author-provided data and code are available. Having such files will usually make the job of replicating a study easier. However, the question of whether having author-provided data and code is good from the perspective of learning is another matter. Most of the time, the analysis of studies published in economics journals was carried out using Stata. If a student obtains these analysis files and translates them to R, it can be a great exercise. However, if the student merely runs author-provided code using author-provided data in the native software used by the study's authors, it can be a good outcome for some students, but there is a real risk that the student won't understand very much of what the program does. This is why I strongly suggest students and researchers seek out the original raw data from IPUMS, and do the replication with these data. Not only will it be a more effective exercise, it will also open up more opportunities for extensions and original research.

Most of the time, journals do not require authors to share their code, and therefore authors do not. In these cases, it can be tempting to email the authors and ask them for their code. In my classes, I explicitly prohibit students from contacting authors, initially. I don't want to encourage students to give up trying on their own too early.

At some point, after trying enough of the likely possible solutions, it is appropriate to contact an author. Having contacted authors a few times during the course of writing this book, I have some advice on how to do this. The importance of "the ask" cannot be overstated. Many authors will not even respond to a request for data and code. Many authors are resistant to the idea of sharing their data and code. Sometimes files get lost or servers get removed, but often authors are worried others will compete with them using their own data, or they don't want their code scrutinized to the point where someone finds a mistake and then says, "gotcha!"

If your aim is to learn from their analysis and not to find errors, try wording your email in a way that indicates this. Making this clear does seem to increase my response rate. Honest mistakes in coding, like in all

walks of life, are inevitable and there's always a collegial way to deal with these situations.[18] If you want authors to share their code with you, it's good to have a reputation as a researcher who is not looking to benefit from errors made by others.

Having described the characteristics of good candidates to replicate, how can you find such studies? If having the author's code is important to you, I suggest starting with those journals that you know have mandatory data and code sharing policies, like the *American Economic Association* journals.[19] Search these journals for the key term "IPUMS." If having the author-supplied code is not important, you'll have a lot more candidates to choose from. You can try searching Google Scholar, and do a cited reference search on Ruggles et al. (2020); all authors who use IPUMS data are supposed to cite this or the earlier versions they used, though not all do. Another great place to search for studies is on the IPUMS-USA webpage, where you will find a bibliography of studies using IPUMS-USA data.

Finally, in the course of writing this book I've picked up some best practices in analyzing the ACS microdata. The top four are:

- Use sample weights. All of the replications illustrate this, using the PERWT and HHWT variables. `script4.R` illustrates sample weights in the context of Chapter 1 review question 6.
- Exclude individuals in group quarters. `script8.R` and `script12.R` (the Orrenius and Zavodny (2015) and Comolli and Bernardi (2015) verifications, respectively) illustrate using the GQ variable.
- Adjust for inflation. `script2.R` (the Winters (2016) verification) illustrates one way to do this; `script9.R` the Chapter 4 code file for software developers, shows another way, using the CPI variable available from IPUMS. Finally `script4.R` illustrates inflation adjustments in the context of Chapter 1 review question 6.

[18] I myself am guilty of publishing an article with an error in it. In Holian (2020b, p. 3) I write, "Electricity expenditures are about 4% smaller in homes built in the 1980s and 1990s in states that adopted building codes, compared to homes built in the 1960s or 1970s in these states." But this is inconsistent with my figure which showed the coefficients on the 1960s and 1970s interaction are about 2.5% higher than the coefficients on the 1980s and 1990s interactions, not 4%. I thank Tue Gørgens for bringing this to my attention. What should an author do when they realize they made an honest error in a published study? Opinions on this differ, but what I did is submit what is called a corrigendum, which the journal published as Holian (2021).

[19] A list of other journals with such policies can be found in Christensen and Miguel (2018).

- Exclude observations with imputed values using data quality flags; `script8.R` illustrates this, using the QBPL and QINCWAGE variables.

Many studies that use the ACS microdata can be improved by better accounting for one or more of the issues in the four bullet points. However, it's not always critical to follow each of them. These are simply best practices, and my goal in highlighting them is to raise awareness among both beginners and professionals.

References

Angrist, Joshua D., and Jörn-Steffen Pischke. Mastering 'metrics: The path from cause to effect. Princeton University Press, 2014.

Angrist, Joshua D., and Jörn-Steffen Pischke. Mostly harmless econometrics: An empiricist's companion. Princeton University Press, 2009.

Bailey, Michael A. Real econometrics: The right tools to answer important questions. Oxford University Press, 2017.

Bailey, James, and Dhaval Dave. "The effect of the Affordable Care Act on entrepreneurship among older adults." Eastern Economic Journal 45, no. 1 (2019): 141–159.

Carilli, Tony. R Companion to Real Econometrics. https://bookdown.org/carillitony/bailey/.

Christensen, Garret, and Edward Miguel. "Transparency, reproducibility, and the credibility of economics research." Journal of Economic Literature 56, no. 3 (2018): 920–980.

Comolli, Chiara Ludovica, and Fabrizio Bernardi. "The causal effect of the great recession on childlessness of white American women." IZA Journal of Labor Economics 4, no. 1 (2015): 1–24.

Costa, Dora L., and Matthew E. Kahn. "Electricity consumption and durable housing: Understanding cohort effects." American Economic Review: Papers & Proceedings 101, no. 3 (2011): 88–92.

Cunningham, Scott. Causal inference: The mixtape. Yale University Press, 2021.

Hanck, Christoph, Martin Arnold, Alexander Gerber, and Martin Schmelzer. "Introduction to econometrics with R." https://www.econometrics-with-r.org/.

Höffler, Jan H. "Teaching replication in quantitative empirical economics." World Economics Association (WEA), Conference on the Economics Curriculum: Towards a Radical Reformation: May 2013.

Holian, Matthew. Data and the American dream. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [research data distributor], March 23, 2021. https://doi.org/10.3886/E135661V1.

Holian, Matthew J. "The impact of urban form on vehicle ownership." Economics Letters 186 (2020a): 108763.

Holian, Matthew J. "The impact of building energy codes on household electricity expenditures." Economics Letters 186 (2020b): 108841.

Holian, Matthew J. "Corrigendum to 'The impact of building energy codes on household electricity expenditures' [Econom. Lett. 186 (2020) 108841]." Economics Letters 200 (2021): 109738.

Janz, Nicole. "Bringing the gold standard into the classroom: Replication in university teaching." International Studies Perspectives 17, no. 4 (2016): 392–407.

Jekel, Marc, Susann Fiedler, Ramona Allstadt Torras, Dorothee Mischkowski, Angela Rachael Dorrough, and Andreas Glöckner. "How to teach open science principles in the undergraduate curriculum—The Hagen Cumulative Science Project." Psychology Learning & Teaching 19, no. 1 (2020): 91–106.

Orrenius, Pia M., and Madeline Zavodny. "The impact of temporary protected status on immigrants' labor market outcomes." American Economic Review: Papers & Proceedings 105, no. 5 (2015): 576–580.

Ruggles, Steven, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas, and Matthew Sobek. IPUMS USA: Version 10.0 [dataset]. Minneapolis, MN: IPUMS, 2020. https://doi.org/10.18128/D010.V10.0.

Stock, James H., and Mark W. Watson. Introduction to econometrics. Boston, MA: Addison Wesley, 2011.

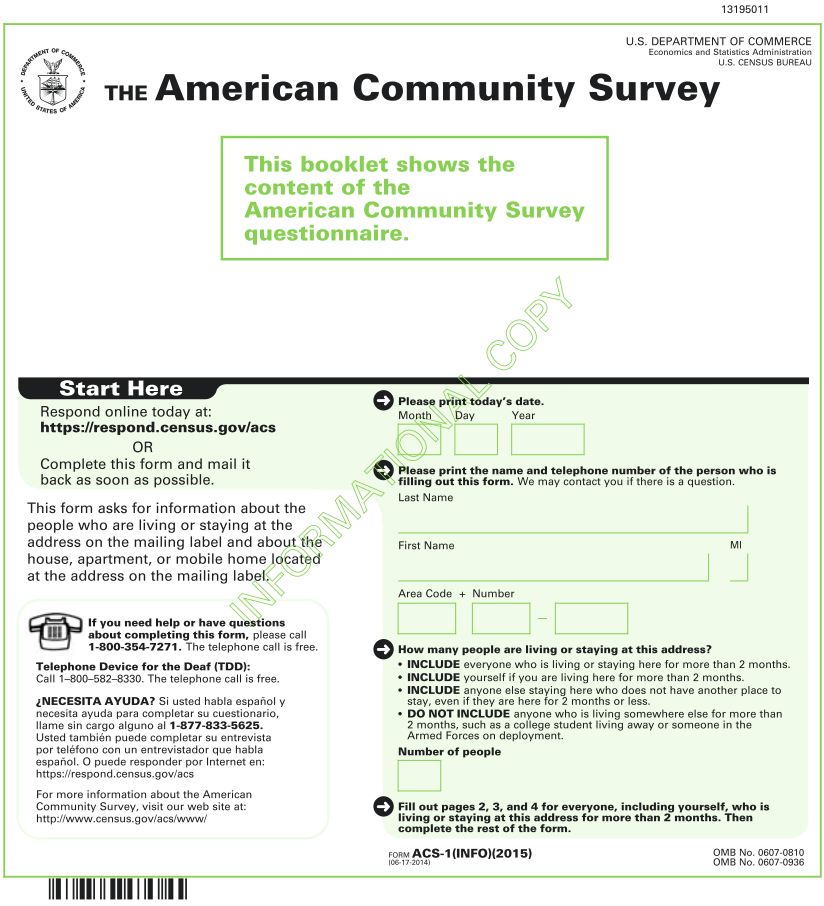
Sundstrom, William A., and Michael J. Kevane. Guide to R: Data analysis for economics. v7.11. https://rpubs.com/wsundstrom/home.

Winters, J. V. "Is economics a good major for future lawyers? Evidence from earnings data." The Journal of Economic Education 47, no. 2 (2016): 187–191.

Appendix B: The ACS Survey Instrument

See Figs. B.1, B.2, B.3, B.3, B.4, B.5, B.6, B.7, B.8, B.9, B.10, B.11, B.12, and B.13.

M. J. Holian, *Data and the American Dream*,
https://doi.org/10.1007/978-3-030-64262-4


13195011

U.S. DEPARTMENT OF COMMERCE
Economics and Statistics Administration
U.S. CENSUS BUREAU

THE American Community Survey

This booklet shows the content of the American Community Survey questionnaire.

Start Here

Respond online today at:
https://respond.census.gov/acs
OR
Complete this form and mail it back as soon as possible.

This form asks for information about the people who are living or staying at the address on the mailing label and about the house, apartment, or mobile home located at the address on the mailing label.

If you need help or have questions about completing this form, please call **1-800-354-7271.** The telephone call is free.

Telephone Device for the Deaf (TDD):
Call 1–800–582–8330. The telephone call is free.

¿NECESITA AYUDA? Si usted habla español y necesita ayuda para completar su cuestionario, llame sin cargo alguno al **1-877-833-5625.** Usted también puede completar su entrevista por teléfono con un entrevistador que habla español. O puede responder por Internet en: https://respond.census.gov/acs

For more information about the American Community Survey, visit our web site at: http://www.census.gov/acs/www/

Please print today's date.
Month Day Year

Please print the name and telephone number of the person who is filling out this form. We may contact you if there is a question.
Last Name
First Name MI
Area Code + Number

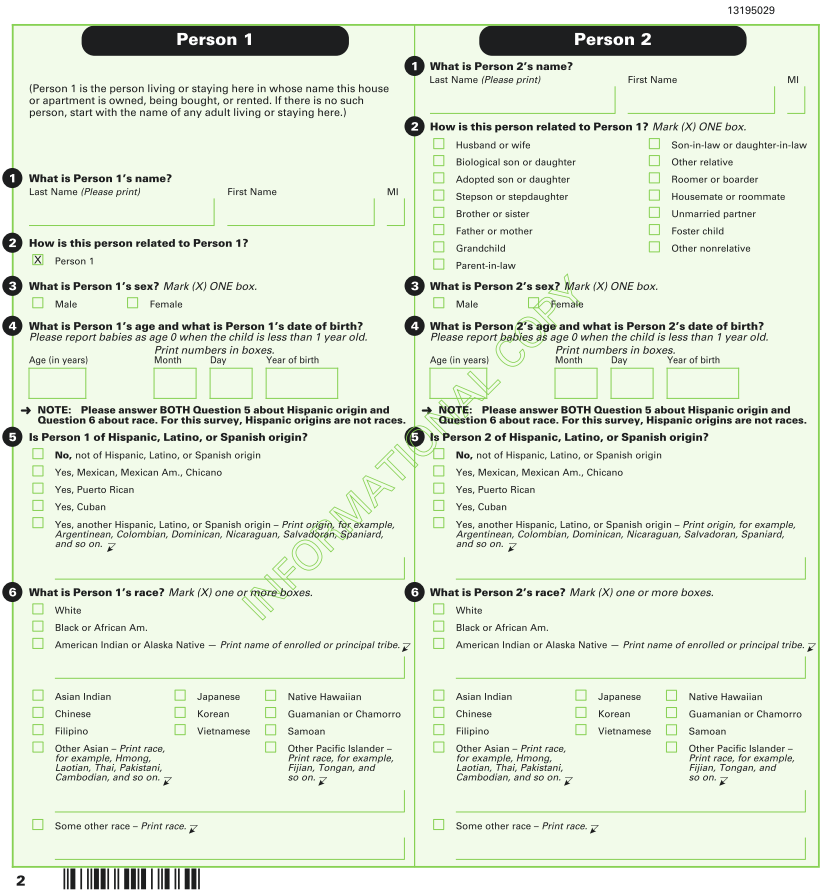
How many people are living or staying at this address?
- **INCLUDE** everyone who is living or staying here for more than 2 months.
- **INCLUDE** yourself if you are living here for more than 2 months.
- **INCLUDE** anyone else staying here who does not have another place to stay, even if they are here for 2 months or less.
- **DO NOT INCLUDE** anyone who is living somewhere else for more than 2 months, such as a college student living away or someone in the Armed Forces on deployment.

Number of people

Fill out pages 2, 3, and 4 for everyone, including yourself, who is living or staying at this address for more than 2 months. Then complete the rest of the form.

FORM **ACS-1(INFO)(2015)** (06-17-2014)
OMB No. 0607-0810
OMB No. 0607-0936

Fig. B.1 Page 1 of the ACS questionnaire



13195029

Person 1

(Person 1 is the person living or staying here in whose name this house or apartment is owned, being bought, or rented. If there is no such person, start with the name of any adult living or staying here.)

1 **What is Person 1's name?**
Last Name *(Please print)* First Name MI

2 **How is this person related to Person 1?**
☒ Person 1

3 **What is Person 1's sex?** *Mark (X) ONE box.*
☐ Male ☐ Female

4 **What is Person 1's age and what is Person 1's date of birth?**
Please report babies as age 0 when the child is less than 1 year old.
Print numbers in boxes.
Age (in years) Month Day Year of birth

→ **NOTE: Please answer BOTH Question 5 about Hispanic origin and Question 6 about race. For this survey, Hispanic origins are not races.**

5 **Is Person 1 of Hispanic, Latino, or Spanish origin?**
☐ **No,** not of Hispanic, Latino, or Spanish origin
☐ Yes, Mexican, Mexican Am., Chicano
☐ Yes, Puerto Rican
☐ Yes, Cuban
☐ Yes, another Hispanic, Latino, or Spanish origin – *Print origin, for example, Argentinean, Colombian, Dominican, Nicaraguan, Salvadoran, Spaniard, and so on.*

6 **What is Person 1's race?** *Mark (X) one or more boxes.*
☐ White
☐ Black or African Am.
☐ American Indian or Alaska Native — *Print name of enrolled or principal tribe.*
☐ Asian Indian ☐ Japanese ☐ Native Hawaiian
☐ Chinese ☐ Korean ☐ Guamanian or Chamorro
☐ Filipino ☐ Vietnamese ☐ Samoan
☐ Other Asian – *Print race, for example, Hmong, Laotian, Thai, Pakistani, Cambodian, and so on.*
☐ Other Pacific Islander – *Print race, for example, Fijian, Tongan, and so on.*
☐ Some other race – *Print race.*

Person 2

1 **What is Person 2's name?**
Last Name *(Please print)* First Name MI

2 **How is this person related to Person 1?** *Mark (X) ONE box.*
☐ Husband or wife
☐ Biological son or daughter
☐ Adopted son or daughter
☐ Stepson or stepdaughter
☐ Brother or sister
☐ Father or mother
☐ Grandchild
☐ Parent-in-law
☐ Son-in-law or daughter-in-law
☐ Other relative
☐ Roomer or boarder
☐ Housemate or roommate
☐ Unmarried partner
☐ Foster child
☐ Other nonrelative

3 **What is Person 2's sex?** *Mark (X) ONE box.*
☐ Male ☐ Female

4 **What is Person 2's age and what is Person 2's date of birth?**
Please report babies as age 0 when the child is less than 1 year old.
Print numbers in boxes.
Age (in years) Month Day Year of birth

→ **NOTE: Please answer BOTH Question 5 about Hispanic origin and Question 6 about race. For this survey, Hispanic origins are not races.**

5 **Is Person 2 of Hispanic, Latino, or Spanish origin?**
☐ **No,** not of Hispanic, Latino, or Spanish origin
☐ Yes, Mexican, Mexican Am., Chicano
☐ Yes, Puerto Rican
☐ Yes, Cuban
☐ Yes, another Hispanic, Latino, or Spanish origin – *Print origin, for example, Argentinean, Colombian, Dominican, Nicaraguan, Salvadoran, Spaniard, and so on.*

6 **What is Person 2's race?** *Mark (X) one or more boxes.*
☐ White
☐ Black or African Am.
☐ American Indian or Alaska Native — *Print name of enrolled or principal tribe.*
☐ Asian Indian ☐ Japanese ☐ Native Hawaiian
☐ Chinese ☐ Korean ☐ Guamanian or Chamorro
☐ Filipino ☐ Vietnamese ☐ Samoan
☐ Other Asian – *Print race, for example, Hmong, Laotian, Thai, Pakistani, Cambodian, and so on.*
☐ Other Pacific Islander – *Print race, for example, Fijian, Tongan, and so on.*
☐ Some other race – *Print race.*

2

Fig. B.2 Page 2 of the ACS questionnaire

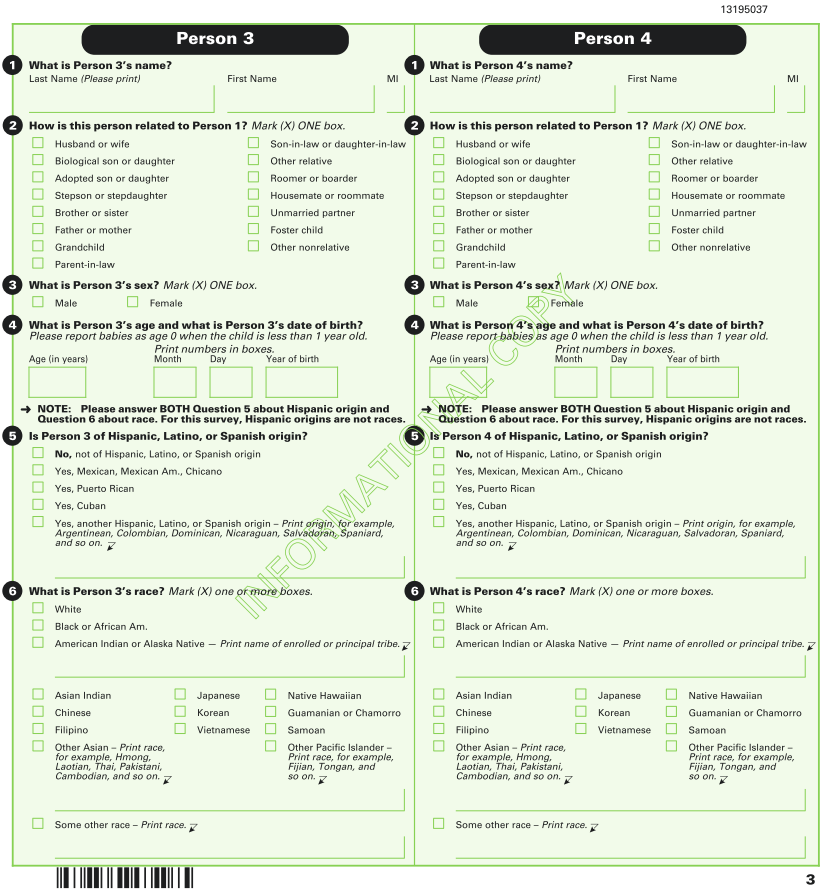
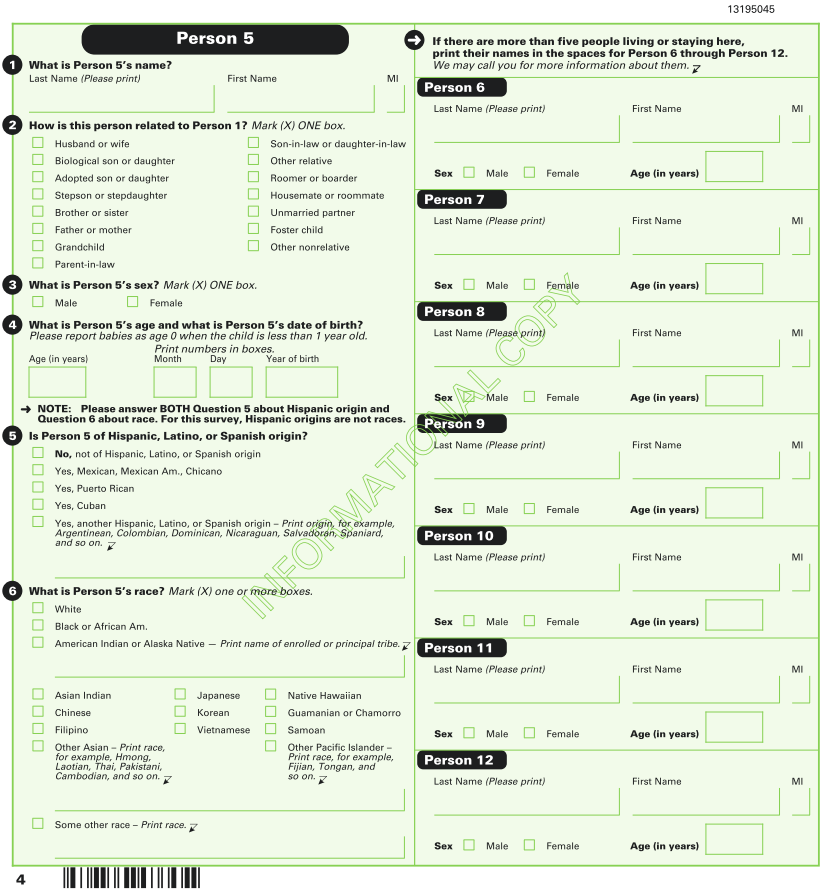


Fig. B.3 Page 3 of the ACS questionnaire



13195045

Person 5

1 What is Person 5's name?
Last Name *(Please print)* | First Name | MI

2 How is this person related to Person 1? *Mark (X) ONE box.*
- ☐ Husband or wife
- ☐ Biological son or daughter
- ☐ Adopted son or daughter
- ☐ Stepson or stepdaughter
- ☐ Brother or sister
- ☐ Father or mother
- ☐ Grandchild
- ☐ Parent-in-law
- ☐ Son-in-law or daughter-in-law
- ☐ Other relative
- ☐ Roomer or boarder
- ☐ Housemate or roommate
- ☐ Unmarried partner
- ☐ Foster child
- ☐ Other nonrelative

3 What is Person 5's sex? *Mark (X) ONE box.*
☐ Male ☐ Female

4 What is Person 5's age and what is Person 5's date of birth?
Please report babies as age 0 when the child is less than 1 year old.
Print numbers in boxes.
Age (in years) | Month | Day | Year of birth

→ **NOTE: Please answer BOTH Question 5 about Hispanic origin and Question 6 about race. For this survey, Hispanic origins are not races.**

5 Is Person 5 of Hispanic, Latino, or Spanish origin?
- ☐ **No,** not of Hispanic, Latino, or Spanish origin
- ☐ Yes, Mexican, Mexican Am., Chicano
- ☐ Yes, Puerto Rican
- ☐ Yes, Cuban
- ☐ Yes, another Hispanic, Latino, or Spanish origin – *Print origin, for example, Argentinean, Colombian, Dominican, Nicaraguan, Salvadoran, Spaniard, and so on.*

6 What is Person 5's race? *Mark (X) one or more boxes.*
- ☐ White
- ☐ Black or African Am.
- ☐ American Indian or Alaska Native — *Print name of enrolled or principal tribe.*
- ☐ Asian Indian
- ☐ Chinese
- ☐ Filipino
- ☐ Other Asian – *Print race, for example, Hmong, Laotian, Thai, Pakistani, Cambodian, and so on.*
- ☐ Japanese
- ☐ Korean
- ☐ Vietnamese
- ☐ Native Hawaiian
- ☐ Guamanian or Chamorro
- ☐ Samoan
- ☐ Other Pacific Islander – *Print race, for example, Fijian, Tongan, and so on.*
- ☐ Some other race – *Print race.*

→ **If there are more than five people living or staying here, print their names in the spaces for Person 6 through Person 12.**
We may call you for more information about them.

Person 6
Last Name *(Please print)* | First Name | MI
Sex ☐ Male ☐ Female **Age (in years)**

Person 7
Last Name *(Please print)* | First Name | MI
Sex ☐ Male ☐ Female **Age (in years)**

Person 8
Last Name *(Please print)* | First Name | MI
Sex ☐ Male ☐ Female **Age (in years)**

Person 9
Last Name *(Please print)* | First Name | MI
Sex ☐ Male ☐ Female **Age (in years)**

Person 10
Last Name *(Please print)* | First Name | MI
Sex ☐ Male ☐ Female **Age (in years)**

Person 11
Last Name *(Please print)* | First Name | MI
Sex ☐ Male ☐ Female **Age (in years)**

Person 12
Last Name *(Please print)* | First Name | MI
Sex ☐ Male ☐ Female **Age (in years)**

4

Fig. B.4 Page 4 of the ACS questionnaire

13195052

Housing

Please answer the following questions about the house, apartment, or mobile home at the address on the mailing label.

1 Which best describes this building? *Include all apartments, flats, etc., even if vacant.*

- ☐ A mobile home
- ☐ A one-family house detached from any other house
- ☐ A one-family house attached to one or more houses
- ☐ A building with 2 apartments
- ☐ A building with 3 or 4 apartments
- ☐ A building with 5 to 9 apartments
- ☐ A building with 10 to 19 apartments
- ☐ A building with 20 to 49 apartments
- ☐ A building with 50 or more apartments
- ☐ Boat, RV, van, etc.

2 About when was this building first built?

- ☐ 2000 or later – *Specify year*
- ☐ 1990 to 1999
- ☐ 1980 to 1989
- ☐ 1970 to 1979
- ☐ 1960 to 1969
- ☐ 1950 to 1959
- ☐ 1940 to 1949
- ☐ 1939 or earlier

3 When did PERSON 1 (listed on page 2) move into this house, apartment, or mobile home?

Month Year

A *Answer questions 4 – 6 if this is a HOUSE OR A MOBILE HOME; otherwise, SKIP to question 7a.*

4 How many acres is this house or mobile home on?

- ☐ Less than 1 acre → *SKIP to question 6*
- ☐ 1 to 9.9 acres
- ☐ 10 or more acres

5 IN THE PAST 12 MONTHS, what were the actual sales of all agricultural products from this property?

- ☐ None
- ☐ $1 to $999
- ☐ $1,000 to $2,499
- ☐ $2,500 to $4,999
- ☐ $5,000 to $9,999
- ☐ $10,000 or more

6 Is there a business (such as a store or barber shop) or a medical office on this property?

- ☐ Yes
- ☐ No

7 a. How many separate rooms are in this house, apartment, or mobile home? *Rooms must be separated by built-in archways or walls that extend out at least 6 inches and go from floor to ceiling.*

- *INCLUDE bedrooms, kitchens, etc.*
- *EXCLUDE bathrooms, porches, balconies, foyers, halls, or unfinished basements.*

Number of rooms

b. How many of these rooms are bedrooms? *Count as bedrooms those rooms you would list if this house, apartment, or mobile home were for sale or rent. If this is an efficiency/studio apartment, print "0".*

Number of bedrooms

8 Does this house, apartment, or mobile home have –

	Yes	No
a. hot and cold running water?	☐	☐
b. a flush toilet?	☐	☐
c. a bathtub or shower?	☐	☐
d. a sink with a faucet?	☐	☐
e. a stove or range?	☐	☐
f. a refrigerator?	☐	☐
g. telephone service from which you can both make and receive calls? *Include cell phones.*	☐	☐

9 At this house, apartment, or mobile home – do you or any member of this household own or use any of the following computers?

- *EXCLUDE GPS devices, digital music players, and devices with only limited computing capabilities, for example: household appliances.*

	Yes	No
a. Desktop, laptop, netbook, or notebook computer	☐	☐
b. Handheld computer, smart mobile phone, or other handheld wireless computer	☐	☐
c. Some other type of computer *Specify*	☐	☐

10 At this house, apartment, or mobile home – do you or any member of this household access the Internet?

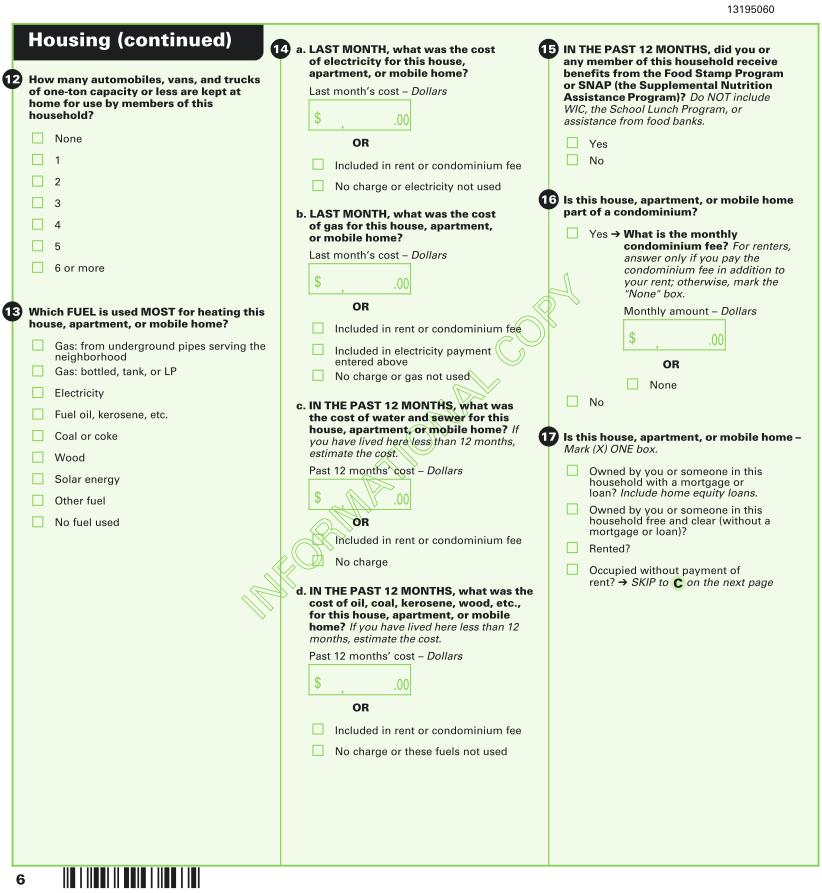
- ☐ Yes, with a subscription to an Internet service
- ☐ Yes, without a subscription to an Internet service → *SKIP to question 12*
- ☐ No Internet access at this house, apartment, or mobile home → *SKIP to question 12*

11 At this house, apartment, or mobile home – do you or any member of this household subscribe to the Internet using –

	Yes	No
a. Dial-up service?	☐	☐
b. DSL service?	☐	☐
c. Cable modem service?	☐	☐
d. Fiber-optic service?	☐	☐
e. Mobile broadband plan for a computer or a cell phone?	☐	☐
f. Satellite Internet service?	☐	☐
g. Some other service? *Specify service*	☐	☐

5

Fig. B.5 Page 5 of the ACS questionnaire



13195060

Housing (continued)

12 **How many automobiles, vans, and trucks of one-ton capacity or less are kept at home for use by members of this household?**

- ☐ None
- ☐ 1
- ☐ 2
- ☐ 3
- ☐ 4
- ☐ 5
- ☐ 6 or more

13 **Which FUEL is used MOST for heating this house, apartment, or mobile home?**

- ☐ Gas: from underground pipes serving the neighborhood
- ☐ Gas: bottled, tank, or LP
- ☐ Electricity
- ☐ Fuel oil, kerosene, etc.
- ☐ Coal or coke
- ☐ Wood
- ☐ Solar energy
- ☐ Other fuel
- ☐ No fuel used

14 **a. LAST MONTH, what was the cost of electricity for this house, apartment, or mobile home?**

Last month's cost – *Dollars*

$, .00

OR

- ☐ Included in rent or condominium fee
- ☐ No charge or electricity not used

b. LAST MONTH, what was the cost of gas for this house, apartment, or mobile home?

Last month's cost – *Dollars*

$, .00

OR

- ☐ Included in rent or condominium fee
- ☐ Included in electricity payment entered above
- ☐ No charge or gas not used

c. IN THE PAST 12 MONTHS, what was the cost of water and sewer for this house, apartment, or mobile home? *If you have lived here less than 12 months, estimate the cost.*

Past 12 months' cost – *Dollars*

$, .00

OR

- ☐ Included in rent or condominium fee
- ☑ No charge

d. IN THE PAST 12 MONTHS, what was the cost of oil, coal, kerosene, wood, etc., for this house, apartment, or mobile home? *If you have lived here less than 12 months, estimate the cost.*

Past 12 months' cost – *Dollars*

$, .00

OR

- ☐ Included in rent or condominium fee
- ☐ No charge or these fuels not used

15 **IN THE PAST 12 MONTHS, did you or any member of this household receive benefits from the Food Stamp Program or SNAP (the Supplemental Nutrition Assistance Program)?** *Do NOT include WIC, the School Lunch Program, or assistance from food banks.*

- ☐ Yes
- ☐ No

16 **Is this house, apartment, or mobile home part of a condominium?**

- ☐ Yes → **What is the monthly condominium fee?** *For renters, answer only if you pay the condominium fee in addition to your rent; otherwise, mark the "None" box.*

  Monthly amount – *Dollars*

  $, .00

  OR

  - ☐ None
- ☐ No

17 **Is this house, apartment, or mobile home –** *Mark (X) ONE box.*

- ☐ Owned by you or someone in this household with a mortgage or loan? *Include home equity loans.*
- ☐ Owned by you or someone in this household free and clear (without a mortgage or loan)?
- ☐ Rented?
- ☐ Occupied without payment of rent? → *SKIP to* **C** *on the next page*

6

Fig. B.6 Page 6 of the ACS questionnaire

13195078

Housing (continued)

B *Answer questions 18a and b if this house, apartment, or mobile home is RENTED. Otherwise, SKIP to question 19.*

18 **a. What is the monthly rent for this house, apartment, or mobile home?**

Monthly amount – *Dollars*

$, .00

b. Does the monthly rent include any meals?

☐ Yes
☐ No

C *Answer questions 19 – 23 if you or any member of this household OWNS or IS BUYING this house, apartment, or mobile home. Otherwise, SKIP to* **E** .

19 **About how much do you think this house and lot, apartment, or mobile home (and lot, if owned) would sell for if it were for sale?**

Amount – *Dollars*

$, , .00

20 **What are the annual real estate taxes on THIS property?**

Annual amount – *Dollars*

$, .00

OR

☐ None

21 **What is the annual payment for fire, hazard, and flood insurance on THIS property?**

Annual amount – *Dollars*

$, .00

OR

☐ None

22 **a. Do you or any member of this household have a mortgage, deed of trust, contract to purchase, or similar debt on THIS property?**

☐ Yes, mortgage, deed of trust, or similar debt
☐ Yes, contract to purchase
☐ No → *SKIP to question 23a*

b. How much is the regular monthly mortgage payment on THIS property?
Include payment only on FIRST mortgage or contract to purchase.

Monthly amount – *Dollars*

$, .00

OR

☐ No regular payment required → *SKIP to question 23a*

c. Does the regular monthly mortgage payment include payments for real estate taxes on THIS property?

☐ Yes, taxes included in mortgage payment
☐ No, taxes paid separately or taxes not required

d. Does the regular monthly mortgage payment include payments for fire, hazard, or flood insurance on THIS property?

☐ Yes, insurance included in mortgage payment
☐ No, insurance paid separately or no insurance

23 **a. Do you or any member of this household have a second mortgage or a home equity loan on THIS property?**

☐ Yes, home equity loan
☐ Yes, second mortgage
☐ Yes, second mortgage and home equity loan
☐ No → *SKIP to* **D**

b. How much is the regular monthly payment on all second or junior mortgages and all home equity loans on THIS property?

Monthly amount – *Dollars*

$, .00

OR

☐ No regular payment required

D *Answer question 24 if this is a MOBILE HOME. Otherwise, SKIP to* **E** .

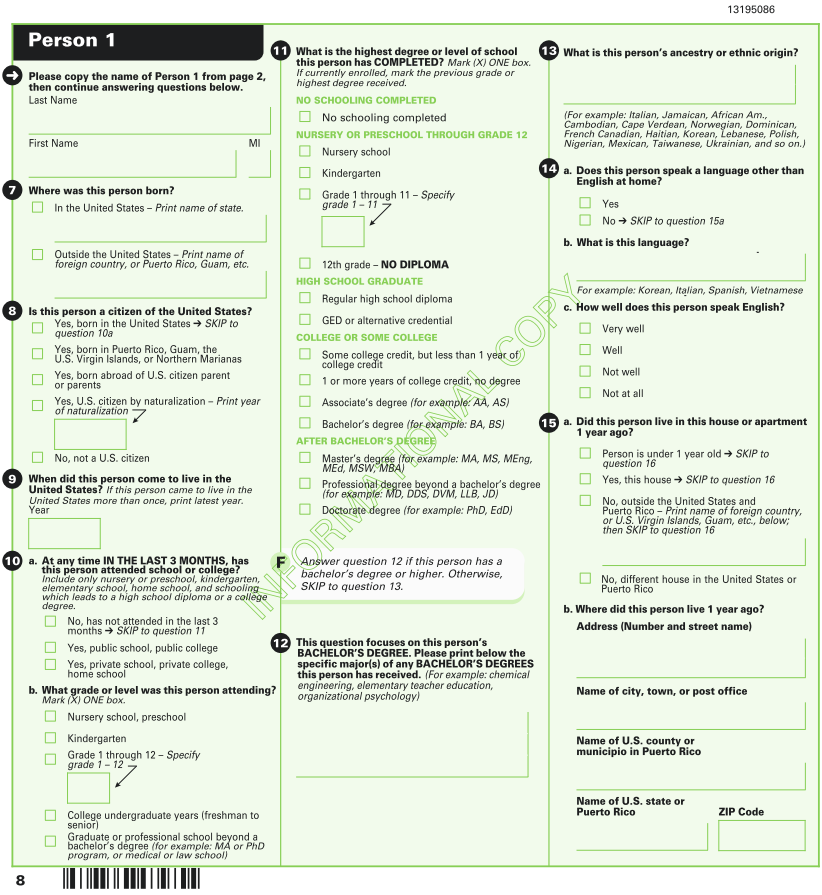
24 **What are the total annual costs for personal property taxes, site rent, registration fees, and license fees on THIS mobile home and its site?**
Exclude real estate taxes.

Annual costs – *Dollars*

$, .00

E *Answer questions about PERSON 1 on the next page if you listed at least one person on page 2. Otherwise, SKIP to page 28 for the mailing instructions.*

Fig. B.7 Page 7 of the ACS questionnaire



13195086

Person 1

Please copy the name of Person 1 from page 2, then continue answering questions below.
Last Name
First Name MI

7 Where was this person born?
- ☐ In the United States – *Print name of state.*
- ☐ Outside the United States – *Print name of foreign country, or Puerto Rico, Guam, etc.*

8 Is this person a citizen of the United States?
- ☐ Yes, born in the United States → *SKIP to question 10a*
- ☐ Yes, born in Puerto Rico, Guam, the U.S. Virgin Islands, or Northern Marianas
- ☐ Yes, born abroad of U.S. citizen parent or parents
- ☐ Yes, U.S. citizen by naturalization – *Print year of naturalization*
- ☐ No, not a U.S. citizen

9 When did this person come to live in the United States? *If this person came to live in the United States more than once, print latest year.*
Year

10 a. At any time IN THE LAST 3 MONTHS, has this person attended school or college? *Include only nursery or preschool, kindergarten, elementary school, home school, and schooling which leads to a high school diploma or a college degree.*
- ☐ No, has not attended in the last 3 months → *SKIP to question 11*
- ☐ Yes, public school, public college
- ☐ Yes, private school, private college, home school

b. What grade or level was this person attending? *Mark (X) ONE box.*
- ☐ Nursery school, preschool
- ☐ Kindergarten
- ☐ Grade 1 through 12 – *Specify grade 1 – 12*
- ☐ College undergraduate years (freshman to senior)
- ☐ Graduate or professional school beyond a bachelor's degree *(for example: MA or PhD program, or medical or law school)*

11 What is the highest degree or level of school this person has COMPLETED? *Mark (X) ONE box. If currently enrolled, mark the previous grade or highest degree received.*

NO SCHOOLING COMPLETED
- ☐ No schooling completed

NURSERY OR PRESCHOOL THROUGH GRADE 12
- ☐ Nursery school
- ☐ Kindergarten
- ☐ Grade 1 through 11 – *Specify grade 1 – 11*
- ☐ 12th grade – **NO DIPLOMA**

HIGH SCHOOL GRADUATE
- ☐ Regular high school diploma
- ☐ GED or alternative credential

COLLEGE OR SOME COLLEGE
- ☐ Some college credit, but less than 1 year of college credit
- ☐ 1 or more years of college credit, no degree
- ☐ Associate's degree *(for example: AA, AS)*
- ☐ Bachelor's degree *(for example: BA, BS)*

AFTER BACHELOR'S DEGREE
- ☐ Master's degree *(for example: MA, MS, MEng, MEd, MSW, MBA)*
- ☐ Professional degree beyond a bachelor's degree *(for example: MD, DDS, DVM, LLB, JD)*
- ☐ Doctorate degree *(for example: PhD, EdD)*

F *Answer question 12 if this person has a bachelor's degree or higher. Otherwise, SKIP to question 13.*

12 This question focuses on this person's BACHELOR'S DEGREE. Please print below the specific major(s) of any BACHELOR'S DEGREES this person has received. *(For example: chemical engineering, elementary teacher education, organizational psychology)*

13 What is this person's ancestry or ethnic origin?
(For example: Italian, Jamaican, African Am., Cambodian, Cape Verdean, Norwegian, Dominican, French Canadian, Haitian, Korean, Lebanese, Polish, Nigerian, Mexican, Taiwanese, Ukrainian, and so on.)

14 a. Does this person speak a language other than English at home?
- ☐ Yes
- ☐ No → *SKIP to question 15a*

b. What is this language?
For example: Korean, Italian, Spanish, Vietnamese

c. How well does this person speak English?
- ☐ Very well
- ☐ Well
- ☐ Not well
- ☐ Not at all

15 a. Did this person live in this house or apartment 1 year ago?
- ☐ Person is under 1 year old → *SKIP to question 16*
- ☐ Yes, this house → *SKIP to question 16*
- ☐ No, outside the United States and Puerto Rico – *Print name of foreign country, or U.S. Virgin Islands, Guam, etc., below; then SKIP to question 16*
- ☐ No, different house in the United States or Puerto Rico

b. Where did this person live 1 year ago?
Address (Number and street name)

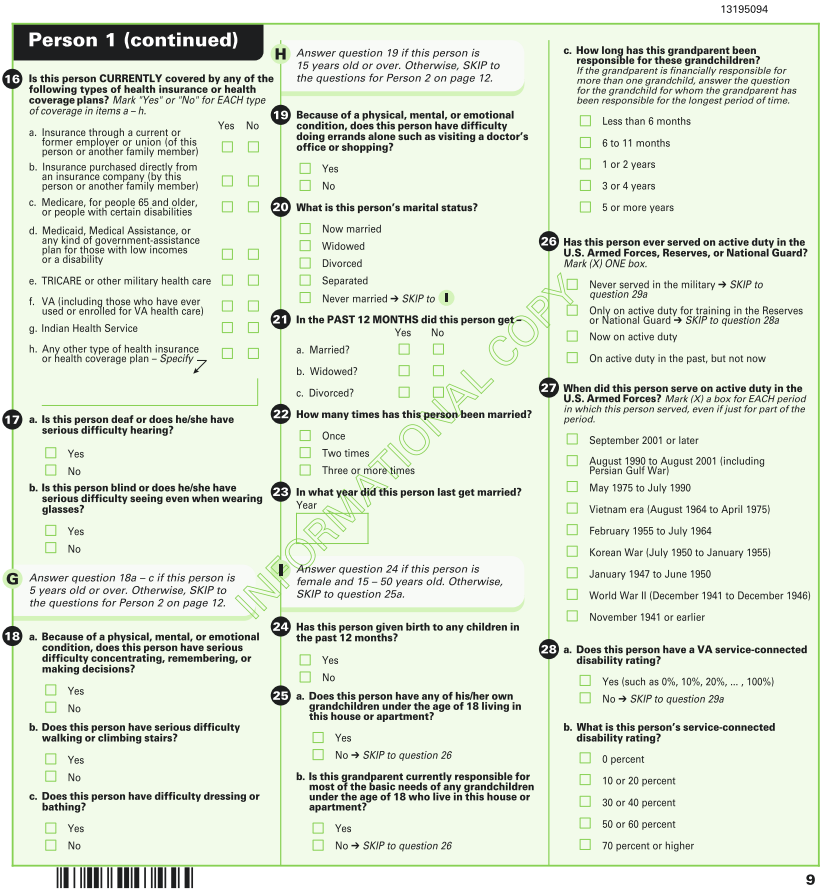
Name of city, town, or post office

Name of U.S. county or municipio in Puerto Rico

Name of U.S. state or Puerto Rico **ZIP Code**

8

Fig. B.8 Page 8 of the ACS questionnaire



13195094

Person 1 (continued)

16 **Is this person CURRENTLY covered by any of the following types of health insurance or health coverage plans?** *Mark "Yes" or "No" for EACH type of coverage in items a – h.*

	Yes	No
a. Insurance through a current or former employer or union (of this person or another family member)	☐	☐
b. Insurance purchased directly from an insurance company (by this person or another family member)	☐	☐
c. Medicare, for people 65 and older, or people with certain disabilities	☐	☐
d. Medicaid, Medical Assistance, or any kind of government-assistance plan for those with low incomes or a disability	☐	☐
e. TRICARE or other military health care	☐	☐
f. VA (including those who have ever used or enrolled for VA health care)	☐	☐
g. Indian Health Service	☐	☐
h. Any other type of health insurance or health coverage plan – *Specify*	☐	☐

17 **a. Is this person deaf or does he/she have serious difficulty hearing?**

☐ Yes
☐ No

b. Is this person blind or does he/she have serious difficulty seeing even when wearing glasses?

☐ Yes
☐ No

G *Answer question 18a – c if this person is 5 years old or over. Otherwise, SKIP to the questions for Person 2 on page 12.*

18 **a. Because of a physical, mental, or emotional condition, does this person have serious difficulty concentrating, remembering, or making decisions?**

☐ Yes
☐ No

b. Does this person have serious difficulty walking or climbing stairs?

☐ Yes
☐ No

c. Does this person have difficulty dressing or bathing?

☐ Yes
☐ No

H *Answer question 19 if this person is 15 years old or over. Otherwise, SKIP to the questions for Person 2 on page 12.*

19 **Because of a physical, mental, or emotional condition, does this person have difficulty doing errands alone such as visiting a doctor's office or shopping?**

☐ Yes
☐ No

20 **What is this person's marital status?**

☐ Now married
☐ Widowed
☐ Divorced
☐ Separated
☐ Never married → *SKIP to* **I**

21 **In the PAST 12 MONTHS did this person get –**

	Yes	No
a. Married?	☐	☐
b. Widowed?	☐	☐
c. Divorced?	☐	☐

22 **How many times has this person been married?**

☐ Once
☐ Two times
☐ Three or more times

23 **In what year did this person last get married?**

Year

I *Answer question 24 if this person is female and 15 – 50 years old. Otherwise, SKIP to question 25a.*

24 **Has this person given birth to any children in the past 12 months?**

☐ Yes
☐ No

25 **a. Does this person have any of his/her own grandchildren under the age of 18 living in this house or apartment?**

☐ Yes
☐ No → *SKIP to question 26*

b. Is this grandparent currently responsible for most of the basic needs of any grandchildren under the age of 18 who live in this house or apartment?

☐ Yes
☐ No → *SKIP to question 26*

c. How long has this grandparent been responsible for these grandchildren? *If the grandparent is financially responsible for more than one grandchild, answer the question for the grandchild for whom the grandparent has been responsible for the longest period of time.*

☐ Less than 6 months
☐ 6 to 11 months
☐ 1 or 2 years
☐ 3 or 4 years
☐ 5 or more years

26 **Has this person ever served on active duty in the U.S. Armed Forces, Reserves, or National Guard?** *Mark (X) ONE box.*

☐ Never served in the military → *SKIP to question 29a*
☐ Only on active duty for training in the Reserves or National Guard → *SKIP to question 28a*
☐ Now on active duty
☐ On active duty in the past, but not now

27 **When did this person serve on active duty in the U.S. Armed Forces?** *Mark (X) a box for EACH period in which this person served, even if just for part of the period.*

☐ September 2001 or later
☐ August 1990 to August 2001 (including Persian Gulf War)
☐ May 1975 to July 1990
☐ Vietnam era (August 1964 to April 1975)
☐ February 1955 to July 1964
☐ Korean War (July 1950 to January 1955)
☐ January 1947 to June 1950
☐ World War II (December 1941 to December 1946)
☐ November 1941 or earlier

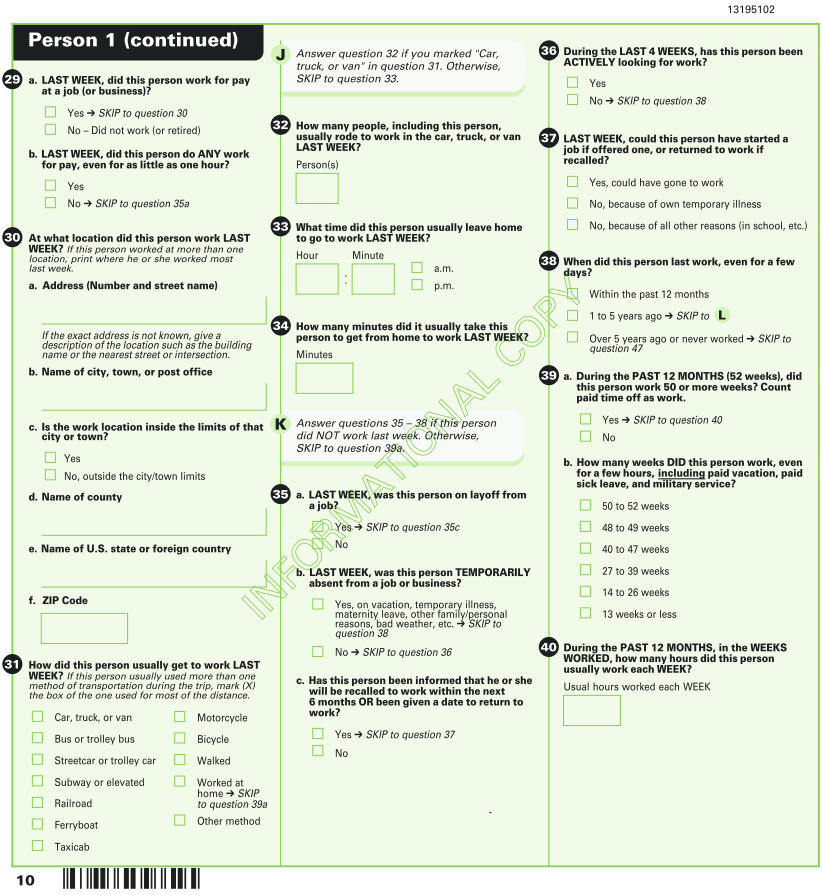
28 **a. Does this person have a VA service-connected disability rating?**

☐ Yes (such as 0%, 10%, 20%, ... , 100%)
☐ No → *SKIP to question 29a*

b. What is this person's service-connected disability rating?

☐ 0 percent
☐ 10 or 20 percent
☐ 30 or 40 percent
☐ 50 or 60 percent
☐ 70 percent or higher

Fig. B.9 Page 9 of the ACS questionnaire


13195102

Person 1 (continued)

29 **a. LAST WEEK, did this person work for pay at a job (or business)?**

- ☐ Yes → *SKIP to question 30*
- ☐ No – Did not work (or retired)

b. LAST WEEK, did this person do ANY work for pay, even for as little as one hour?

- ☐ Yes
- ☐ No → *SKIP to question 35a*

30 **At what location did this person work LAST WEEK?** *If this person worked at more than one location, print where he or she worked most last week.*

a. Address (Number and street name)

If the exact address is not known, give a description of the location such as the building name or the nearest street or intersection.

b. Name of city, town, or post office

c. Is the work location inside the limits of that city or town?

- ☐ Yes
- ☐ No, outside the city/town limits

d. Name of county

e. Name of U.S. state or foreign country

f. ZIP Code

31 **How did this person usually get to work LAST WEEK?** *If this person usually used more than one method of transportation during the trip, mark (X) the box of the one used for most of the distance.*

- ☐ Car, truck, or van
- ☐ Bus or trolley bus
- ☐ Streetcar or trolley car
- ☐ Subway or elevated
- ☐ Railroad
- ☐ Ferryboat
- ☐ Taxicab
- ☐ Motorcycle
- ☐ Bicycle
- ☐ Walked
- ☐ Worked at home → *SKIP to question 39a*
- ☐ Other method

J *Answer question 32 if you marked "Car, truck, or van" in question 31. Otherwise, SKIP to question 33.*

32 **How many people, including this person, usually rode to work in the car, truck, or van LAST WEEK?**

Person(s)

33 **What time did this person usually leave home to go to work LAST WEEK?**

Hour : Minute

- ☐ a.m.
- ☐ p.m.

34 **How many minutes did it usually take this person to get from home to work LAST WEEK?**

Minutes

K *Answer questions 35 – 38 if this person did NOT work last week. Otherwise, SKIP to question 39a.*

35 **a. LAST WEEK, was this person on layoff from a job?**

- ☐ Yes → *SKIP to question 35c*
- ☐ No

b. LAST WEEK, was this person TEMPORARILY absent from a job or business?

- ☐ Yes, on vacation, temporary illness, maternity leave, other family/personal reasons, bad weather, etc. → *SKIP to question 38*
- ☐ No → *SKIP to question 36*

c. Has this person been informed that he or she will be recalled to work within the next 6 months OR been given a date to return to work?

- ☐ Yes → *SKIP to question 37*
- ☐ No

36 **During the LAST 4 WEEKS, has this person been ACTIVELY looking for work?**

- ☐ Yes
- ☐ No → *SKIP to question 38*

37 **LAST WEEK, could this person have started a job if offered one, or returned to work if recalled?**

- ☐ Yes, could have gone to work
- ☐ No, because of own temporary illness
- ☐ No, because of all other reasons (in school, etc.)

38 **When did this person last work, even for a few days?**

- ☐ Within the past 12 months
- ☐ 1 to 5 years ago → *SKIP to* **L**
- ☐ Over 5 years ago or never worked → *SKIP to question 47*

39 **a. During the PAST 12 MONTHS (52 weeks), did this person work 50 or more weeks? Count paid time off as work.**

- ☐ Yes → *SKIP to question 40*
- ☐ No

b. How many weeks DID this person work, even for a few hours, including paid vacation, paid sick leave, and military service?

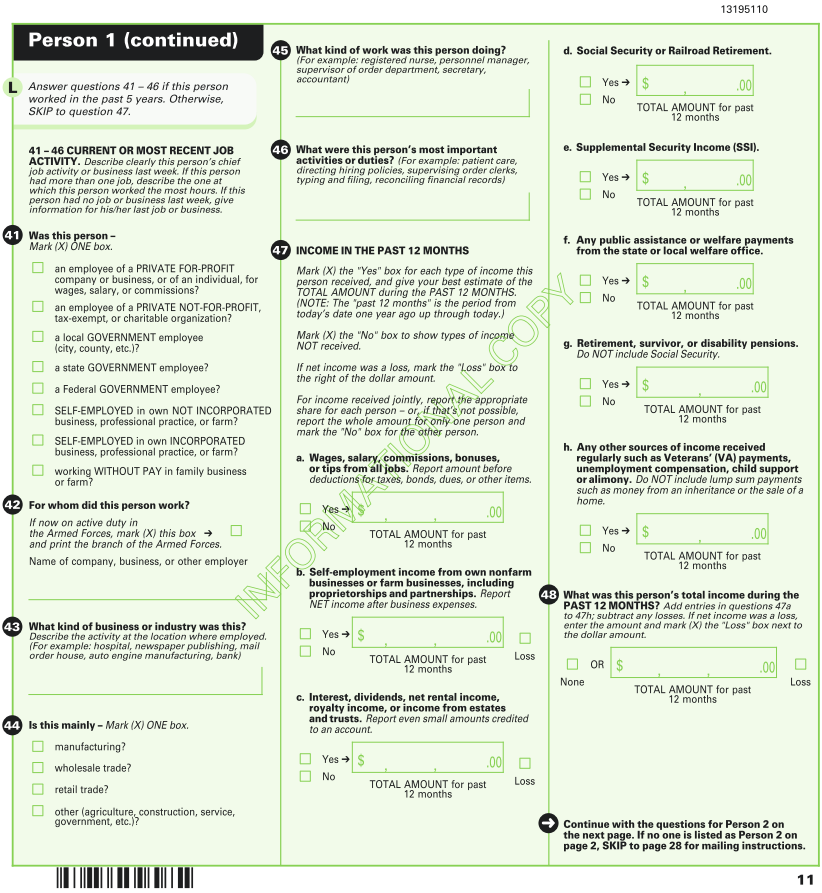
- ☐ 50 to 52 weeks
- ☐ 48 to 49 weeks
- ☐ 40 to 47 weeks
- ☐ 27 to 39 weeks
- ☐ 14 to 26 weeks
- ☐ 13 weeks or less

40 **During the PAST 12 MONTHS, in the WEEKS WORKED, how many hours did this person usually work each WEEK?**

Usual hours worked each WEEK

10

Fig. B.10 Page 10 of the ACS questionnaire



13195110

Person 1 (continued)

L *Answer questions 41 – 46 if this person worked in the past 5 years. Otherwise, SKIP to question 47.*

41 – 46 CURRENT OR MOST RECENT JOB ACTIVITY. *Describe clearly this person's chief job activity or business last week. If this person had more than one job, describe the one at which this person worked the most hours. If this person had no job or business last week, give information for his/her last job or business.*

41 Was this person –
Mark (X) ONE box.

- ☐ an employee of a PRIVATE FOR-PROFIT company or business, or of an individual, for wages, salary, or commissions?
- ☐ an employee of a PRIVATE NOT-FOR-PROFIT, tax-exempt, or charitable organization?
- ☐ a local GOVERNMENT employee (city, county, etc.)?
- ☐ a state GOVERNMENT employee?
- ☐ a Federal GOVERNMENT employee?
- ☐ SELF-EMPLOYED in own NOT INCORPORATED business, professional practice, or farm?
- ☐ SELF-EMPLOYED in own INCORPORATED business, professional practice, or farm?
- ☐ working WITHOUT PAY in family business or farm?

42 For whom did this person work?
If now on active duty in the Armed Forces, mark (X) this box → ☐ *and print the branch of the Armed Forces.*

Name of company, business, or other employer

43 What kind of business or industry was this?
Describe the activity at the location where employed. (For example: hospital, newspaper publishing, mail order house, auto engine manufacturing, bank)

44 Is this mainly – *Mark (X) ONE box.*

- ☐ manufacturing?
- ☐ wholesale trade?
- ☐ retail trade?
- ☐ other (agriculture, construction, service, government, etc.)?

45 What kind of work was this person doing?
(For example: registered nurse, personnel manager, supervisor of order department, secretary, accountant)

46 What were this person's most important activities or duties? *(For example: patient care, directing hiring policies, supervising order clerks, typing and filing, reconciling financial records)*

47 INCOME IN THE PAST 12 MONTHS

Mark (X) the "Yes" box for each type of income this person received, and give your best estimate of the TOTAL AMOUNT during the PAST 12 MONTHS. (NOTE: The "past 12 months" is the period from today's date one year ago up through today.)

Mark (X) the "No" box to show types of income NOT received.

If net income was a loss, mark the "Loss" box to the right of the dollar amount.

For income received jointly, report the appropriate share for each person – or, if that's not possible, report the whole amount for only one person and mark the "No" box for the other person.

a. Wages, salary, commissions, bonuses, or tips from all jobs. *Report amount before deductions for taxes, bonds, dues, or other items.*

☐ Yes → $ .00
☐ No
TOTAL AMOUNT for past 12 months

b. Self-employment income from own nonfarm businesses or farm businesses, including proprietorships and partnerships. *Report NET income after business expenses.*

☐ Yes → $ .00 ☐ Loss
☐ No
TOTAL AMOUNT for past 12 months

c. Interest, dividends, net rental income, royalty income, or income from estates and trusts. *Report even small amounts credited to an account.*

☐ Yes → $ .00 ☐ Loss
☐ No
TOTAL AMOUNT for past 12 months

d. Social Security or Railroad Retirement.

☐ Yes → $ .00
☐ No
TOTAL AMOUNT for past 12 months

e. Supplemental Security Income (SSI).

☐ Yes → $ .00
☐ No
TOTAL AMOUNT for past 12 months

f. Any public assistance or welfare payments from the state or local welfare office.

☐ Yes → $ .00
☐ No
TOTAL AMOUNT for past 12 months

g. Retirement, survivor, or disability pensions. *Do NOT include Social Security.*

☐ Yes → $ .00
☐ No
TOTAL AMOUNT for past 12 months

h. Any other sources of income received regularly such as Veterans' (VA) payments, unemployment compensation, child support or alimony. *Do NOT include lump sum payments such as money from an inheritance or the sale of a home.*

☐ Yes → $ .00
☐ No
TOTAL AMOUNT for past 12 months

48 What was this person's total income during the PAST 12 MONTHS? *Add entries in questions 47a to 47h; subtract any losses. If net income was a loss, enter the amount and mark (X) the "Loss" box next to the dollar amount.*

☐ None OR $ .00 ☐ Loss
TOTAL AMOUNT for past 12 months

Continue with the questions for Person 2 on the next page. If no one is listed as Person 2 on page 2, SKIP to page 28 for mailing instructions.

Fig. B.11 Page 11 of the ACS questionnaire

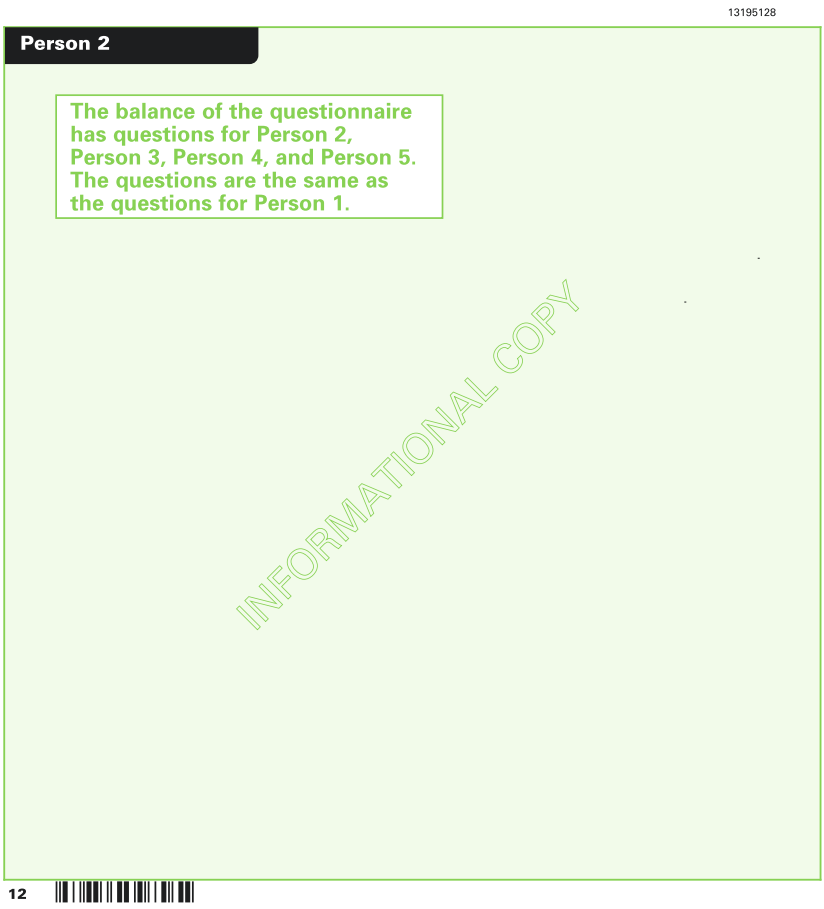
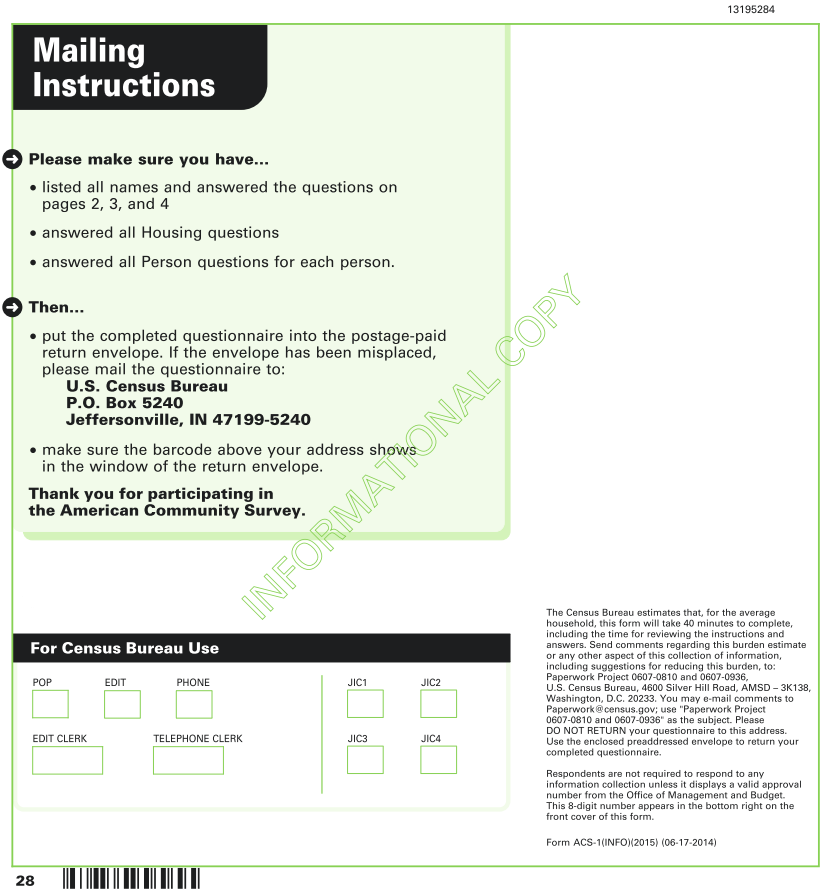


Fig. B.12 Page 12 of the ACS questionnaire



13195284

Mailing Instructions

Please make sure you have...

- listed all names and answered the questions on pages 2, 3, and 4
- answered all Housing questions
- answered all Person questions for each person.

Then...

- put the completed questionnaire into the postage-paid return envelope. If the envelope has been misplaced, please mail the questionnaire to:
  U.S. Census Bureau
  P.O. Box 5240
  Jeffersonville, IN 47199-5240
- make sure the barcode above your address shows in the window of the return envelope.

Thank you for participating in the American Community Survey.

For Census Bureau Use

POP | EDIT | PHONE | JIC1 | JIC2

EDIT CLERK | TELEPHONE CLERK | JIC3 | JIC4

The Census Bureau estimates that, for the average household, this form will take 40 minutes to complete, including the time for reviewing the instructions and answers. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to: Paperwork Project 0607-0810 and 0607-0936, U.S. Census Bureau, 4600 Silver Hill Road, AMSD – 3K138, Washington, D.C. 20233. You may e-mail comments to Paperwork@census.gov; use "Paperwork Project 0607-0810 and 0607-0936" as the subject. Please DO NOT RETURN your questionnaire to this address. Use the enclosed preaddressed envelope to return your completed questionnaire.

Respondents are not required to respond to any information collection unless it displays a valid approval number from the Office of Management and Budget. This 8-digit number appears in the bottom right on the front cover of this form.

Form ACS-1(INFO)(2015) (06-17-2014)

28

Fig. B.13 The last page of the ACS questionnaire

Glossary

average causal effect Individuals respond differently to treatment; the average of the individual causal effects is the average causal effect.

bad control An independent variable that is an outcome related to the variable of interest; although it may satisfy the two OVB conditions, it should not be used as a control because it does not unambiguously reduce bias in the estimate of the coefficient on the variable of interest. See also post-treatment variable.

basic D-in-D model A multivariate regression model with three independent variables: (1) a treatment group indicator, (2) an indicator of when treatment was in effect for the treatment group, and (3) the interaction between (1) and (2). The interaction is the independent variable of interest.

basic D-in-D with control variables A basic D-in-D model, which contains three independent variables, plus additional right-hand side variables that are included to reduce bias in the estimate of the coefficient on the interaction term.

best-case scenario In CBA, a set of assumptions that results in the highest possible NPV calculation. See worst-case scenario.

big data analytics A term that refers to analyzing large data sets, such as data on web transactions, administrative data, or public use microdata.

M. J. Holian, *Data and the American Dream*,
https://doi.org/10.1007/978-3-030-64262-4

binary variable A variable that takes on two values, 0 and 1. For example, when a person's gender is recorded as either male (0) or female (1). Also called an indicator or dummy variable.

bivariate regression A regression model that uses only two variables, one dependent and one independent.

categorical variable A variable, like STATEFIP, with multiple categories but no inherent ordering.

causal inference The art and science of drawing conclusions about cause and effect relationships. We often see estimates of average causal effects.

codebook A document or file that tells us the precise meaning of all possible values for all variables in a dataset.

coefficients See regression coefficients.

constant Also called an intercept; in a bivariate regression equation, this is the expected value of the dependent variable when the independent variable is zero.

control variable A variable that is included in a multivariate regression model in order to reduce bias in the estimate of the coefficient on the main independent variable of interest. See Regression Control.

cost-benefit analysis (CBA) A decision-making tool that expresses all positive and negative impacts of a policy or project, to all members of society, in present value monetary terms, and recommends the course of action with the highest net present value.

dependent variable A variable that measures the outcome of interest. Usually denoted with a Y. Also called a left-hand side variable.

descriptive statistic A statistic, such as a mean or proportion, that summarizes a characteristic of a sample or population, but does not indicate how or why the characteristics are how they are.

difference-in-differences (D-in-D) A causal inference technique that relies on a natural experiment to identify a treatment group in observational data. Various models are used in estimation: basis D-in-D, basic D-in-D with control variables, fixed effect D-in-D, and TWFE.

difference in means The difference in the mean value of some variable across two groups. It can be calculated by subtracting one group average from the other, or by estimating a bivariate regression model with a binary independent variable.

economic analysis Although this term sounds so broad that it could apply to almost anything economists do, it is often reserved for economic analysis of decision-making. Examples include CBA, EIA, and FIA.

Economic Impact Analysis (EIA) A specific type of economic analysis that focuses on spending. The definition of standing is often narrow in EIA; for example, an EIA may only measure the impact of a new sports stadium on spending in a city, and not distinguish between brand new spending and spending that would have otherwise occurred in an adjacent city.

endogeneity A situation in a regression model where the error term is correlated with the main independent variable of interest. Several situations give rise to endogeneity, including OVB, sample selection bias, and failure to use logged variables or polynomial models when they are appropraite.

estimation subsample A subset of the full set of raw data that is used to estimate a model or statistic. For example, the complete ACS data is representative of all Americans, but a study may estimate statistics using a subset of homeowners from California.

error term In a regression model, a term that represents unmeasured factors. Mathematically, it is the difference between the value of Y and the population regression function, but because we never observe the population regression function, we never observe the value of the error term. We do observe the difference between Y and the value predicted by the sample regression function, but this difference is the residual, not the error term.

exogeneity In an instrumental variables model, an instrument satisfies the exogeneity condition if it does not affect the dependent variable except through its impact on the main independent variable of interest. It is generally not possible to test whether an instrument is exogenous or not and is usually established based on logical argument. See valid instrument.

experimental data Data generated through a randomized, controlled experiment.

extension Estimating a model from the literature using data on a different population.

falsification test In D-in-D methodology, if treatment effects are non-zero using an alternative control group as a treatment group, it casts doubt on D-in-D estimates.

fiscal impact analysis (FIA) A specific type of economic analysis that focuses on government spending. For example, an FIA of housing the homeless may focus on the costs of providing housing versus the costs of providing police, ambulance, court, and other public services in the absence of the housing. The happiness of the housed or of those in the neighborhoods where the formally homeless lived would not be taken into account in an FIA.

fitted values A predicted value of the dependent variable. To find them using an estimated equation, plug-in researcher-specified values of the independent variable (usually the actual values of them for an observation) and recover the model's prediction.

fixed effects In a regression model, these are binary variables included to represent all but one category for a categorical variable (the effect in the excluded category is reflected in the constant term). For example, including state fixed effects in a model means we include binary variables for all but one of the states represented in the estimation subsample. As another example, including survey year fixed effects means we include binary variables for all but one of the survey years.

first-stage equation A regression equation where the main independent variable of interest is on the left-hand side, and the instrument (and potentially other variables) are on the right. See instrumental variables and second-stage equation.

fixed effect D-in-D model Like the basic D-in-D, this model includes a single interaction term which is the product of a treatment group indicator and a treatment period indicator, but unlike the basic D-in-D, it also includes a full set of periods and group indicators, not just two.

ideal experiment When a researcher imagines a world without ethical or financial constraints, and considers what experiment could estimate the causal effect of interest. Describing an ideal experiment defines the causal effect one aims to measure with observational data.

independent variable A variable that is used to explain or predict the dependent variable. Usually denoted with X. Also called right-hand side variable.

independent variable of interest An independent variable that is of primary interest in a regression control study. It is not a control variable.

inferential statistics The branch of the field of statistics that seeks to estimate population characteristics using samples.

instrument See instrumental variables.

instrumental variables A causal inference technique that requires having a variable (the instrument) that predicts the main independent variable of interest, but that does not otherwise determine the dependent variable.

interaction model A regression model where one of the right-hand side variables is the product of two of the other right-hand side variables. The basic difference-in-differences model uses an interaction model.

linear probability model A bivariate or multivariate regression model where the dependent variable is binary. Fitted values are interpreted as predicted probabilities.

literature When academics refer to "the literature" they are referring to all academic journal articles and books that have been published on a specific topic.

logged variable A dependent or independent variable that has been transformed using the natural logarithmic function.

mean A simple average. Calculated by summing up the values of a variable for all observations, and dividing by the number of observations.

merged data Data a researcher mergers on to the main estimation subsample. For example, the average January temperature of the respondent's PUMA is not available in the IPUMS-USA data and must be obtained from another source and merged on by the researcher. Complete replication files include all analysis files, raw data files, and all merged data.

microdata Data where the unit of observation is a person or household. This is in contrast to aggregate data where the unit of observation may be a PUMA, city or state.

multivariate regression A regression model that uses more than one independent variable. Both the regression control and D-in-D techniques use multivariate regression models.

natural experiment A setting identified by a researcher where natural, social, political, or other processes end up assigning treatment in a way that

is as if it were randomly assigned by an experimenter in a true experiment, even though it was not.

net present value (NPV) Future dollars are not worth as much as current dollars. To convert a future value to present value we must discount it. The NPV of a social project is the present value of its benefits minus the present value of its costs.

observational data Data generated through observing or measuring behavior, such as through a survey, that is not experimental.

omitted variables bias (OVB) When the coefficient on the main independent variable of interest X systematically overstates or understates the true treatment effect. It occurs by failing to include a control variable that meets two conditions: the omitted control (1) explains Y and (2) is correlated with X.

ordinal variable A variable, like EDUCD, with multiple categories and an intrinsic ordering, but that is not numerical.

ordinary least squares A mathematical method for estimating regression coefficients which minimizes the sum of squared residuals.

original research An analysis may be inspired by a previously published study, but if it estimates a different model, on a different population, it is not a replication, reanalysis, or extension, but rather just original research.

perfect multicolinearity A setting in which one independent variable is perfectly explained by another independent variable (such as in a model that includes both MALE and FEMALE on the right-hand side), or, is perfectly explained by a linear combination of the other independent variables (such as in a model with a constant term, where one of the fixed effects is not excluded).

polynomial model A multivariate regression model where an independent variable is included both in levels, as well as in squared (or higher order) terms. An example is including both AGE and the square of AGE on the right-hand side.

post-treatment variable See bad control.

pre-trends analysis In D-in-D methodology, if pre-treatment period trends move in parallel for treatment and control groups, it strengthens the case for interpreting the D-in-D estimate as resulting from a natural experiment.

public use microdata Publicly-available, individual-response data, usually collected by governmental Census bureaus. Identifying variables (the respondent's name and address) are removed to preserve confidentiality.

randomized experiment Uses random assignment to place subjects into control and treatment group. After administering the treatment to the treatment group, it typically compares average outcomes for the two groups. Any difference in outcome is presumably only due to the fact that one group received the treatment, as randomization should have ensured the two groups, prior to the experiment, had average characteristics that were essentially identical.

random sampling Surveys that aim to be representative of a population often randomly sample the population. However, differences in response rates and other factors can lead to non-representative samples. In these cases, sampling weights are often used to make estimates more representative. Contrast with randomized experiment.

reanalysis Using the same sample of data as a previously published study to estimate different models.

regression - A technique for estimating linear equations to measure empirical relationships. Sometimes called OLS after the mathematical technique that underlies regression.

regression coefficients The parameters of the regression equation, usually denoted with Greek letters. When these parameters are estimated with data they are statistics. Bivariate regression equations have one constant and one slope coefficient, and multivariate regression equations have one constant and multiple slope coefficients.

regression control A causal inference technique that uses a multivariate regression model and data on control variables to eliminate or reduce omitted variables bias in the coefficient on the main variable of interest.

relevant instrument In an instrumental variables model, an instrument that predicts the main independent variable of interest. See valid instrument.

repeated cross-section A survey conducted in one year is a cross-sectional data set. If the same survey is conducted each year on a different sample of individuals, it is a repeated cross-sectional data set, whereas if it is conducted each year on the same sample it is called a panel or longitudinal data set.

replicate and extend An approach to empirical research that starts with verifying previously published estimates of a model (a verification), and then carries out a replication, reproduction, reanalysis, extension, or original research inspired by the original study.

replication Both verifications and reproductions are considered replications.

reproduction Estimating a model from the literature on an identical population but using a different sample.

residual The difference between the actual observed value of the dependent variable, and the fitted value. Contrast with the error term.

sample weights When survey sampling is known not to be representative of the population based on certain characteristics, the data distributor will usually provide sample weights, which can be used with software programs to transform the data which ideally make it more representative of the population the sample was drawn from. See random sampling.

sample selection bias When data are obtained through non-representative sampling. See random sampling. Compare with selection bias.

sampling variation Because subjects in a population are randomly selected for sampling, estimates of population characteristics will vary from one sample to another, even if the characteristics of the population remain constant.

second-stage equation The instrumental variables technique is usually carried out using the two-stage least squares model. The second-stage part of the equation uses predicted values from the first-stage as the independent variable of interest.

selection bias When a difference in means calculated with observational data systematically overstates or understates the average causal effect. The gap between the difference in means and average causal effect is referred to as selection bias, self-selection bias or the selection effect. Compare with sample selection bias.

selection effect See selection bias.

shadow price When market prices do not reflect social costs or benefits, economists estimate shadow prices, which are simply the true social value of an impact.

statistical inference The process of drawing conclusions about populations from samples of data. If a relationship observed in a sample is likely to hold up in repeated samples, it is said to be statistically significant.

statistics Estimates of population parameters that are calculated with data. Means, medians, and regression coefficients are all statistics. This term also refers to a field of mathematics, but this is not how it is used in this book.

standard error The "standard error of a regression coefficient" measures sampling variance in regression coefficients. If this standard error is small relative to the estimated regression coefficient, the estimate is said to be statistically significant. Another type of standard error, "the standard error of the regression" is used as a goodness of fit measure for the regression model, not to determine statistical significance; this type of standard error, often reported by statistical software programs as default regression output, is the standard deviation of the residual, and a small value indicates the model predictions are usually close to the actual Y values.

standing In CBA, the definition of society. Whoever's preferences count has standing.

time horizon In CBA, the length of time the policy or project is assumed to generate benefits or costs.

top code Data distributors often do not reveal the highest responses to questions like income to preserve the confidentiality of survey respondents, and to insure data quality issues (such as misreporting or clerical error) do not bias estimates.

treatment effect In a laboratory, ideal or natural experiment, the effect of the treatment on the outcome. Also called the causal effect.

two-way fixed effect (TWFE) estimator A model for repeated cross-sectional data that includes fixed effects at the unit (person or household) and time levels. It is sometimes considered a type of difference-in-differences model.

valid instrument An instrument that is relevant and exogenous.

value of a statistical life An estimate of the sum of what individuals in society are willing to pay to reduce the probability of a fatality by one.

variable. Data on an outcome or characteristic of the person, household, or other units of observation under study.

variable of interest See independent variable of interest.

verification Estimating the same model, with the same sample of data used by the authors of the original study.

worst-case scenario In CBA, a set of assumptions that results in the lowest possible NPV calculation. See best-case scenario.

References

Abraham, Katharine G., John Haltiwanger, Kristin Sandusky, and James Spletzer. "The rise of the gig economy: Fact or fiction?" AEA Papers and Proceedings 109 (2019): 357–361.

Allen, Treb, Cauê de Castro Dobbin, and Melanie Morten. "Border walls." No. w25267. National Bureau of Economic Research, February 2019.

Amuedo-Dorantes, Catalina, Esther Arenas-Arroyo, and Almudena Sevilla. "Immigration enforcement and economic resources of children with likely unauthorized parents." Journal of Public Economics 158 (2018): 63–78.

Angrist, Joshua D., and Jörn-Steffen Pischke. Mastering 'metrics: The path from cause to effect. Princeton University Press, 2014.

Angrist, Joshua D., and Jörn-Steffen Pischke. Mostly harmless econometrics: An empiricist's companion. Princeton University Press, 2009.

Argys, Laura M., and Susan L. Averett. "The effect of family size on education: New evidence from China's one-child policy." Journal of Demographic Economics 85, no. 1 (2019): 21–42.

Aroonruengsawat, Anin, Maximilian Auffhammer, and Alan H. Sanstad. "The impact of state level building codes on residential electricity consumption." Energy Journal-Cleveland 33, no. 1 (2012): 31.

Bailey, Michael A. Real econometrics: The right tools to answer important questions. Oxford University Press, 2017.

M. J. Holian, *Data and the American Dream*,
https://doi.org/10.1007/978-3-030-64262-4

Bailey, James, and Dhaval Dave. "The effect of the Affordable Care Act on entrepreneurship among older adults." Eastern Economic Journal 45, no. 1 (2019): 141–159.

Bailey, James. "Health insurance and the supply of entrepreneurs: New evidence from the affordable care act." Small Business Economics 49, no. 3 (2017): 627–646.

Blau, F. D., L. M. Kahn, P. Brummund, J. Cook, and M. Larson-Koester. "Is there still Son preference in the United States?" Journal of Population Economics 33, no. 3 (2020): 709–750.

Bleemer, Zachary, and Aashish Mehta. "Will studying economics make you rich? A regression discontinuity analysis of the returns to college major." American Economic Journal: Applied Economics. Forthcoming, 2021.

Boardman, Anthony E., David H. Greenberg, Aidan R. Vining, and David L. Weimer. Cost-benefit analysis: Concepts and practice. Cambridge University Press, 2017. https://doi.org/10.17226/23550.

Callaway, Brantly, and Pedro H.C. Sant'Anna. "Difference-in-differences with multiple time periods." Journal of Econometrics (in press, 2020). https://doi.org/10.1016/j.jeconom.2020.12.001.

Census Bureau. "Demographic trends in the 20th century." Census 2000 Special Reports, 2002. https://www.census.gov/prod/2002pubs/censr-4.pdf

Chaudhuri, Anoshua, and Susan G. Zieff. "Do open streets initiatives impact local businesses? The case of Sunday Streets in San Francisco, California." Journal of Transport & Health 2, no. 4 (2015): 529–539.

Christensen, Garret, and Edward Miguel. "Transparency, reproducibility, and the credibility of economics research." Journal of Economic Literature 56, no. 3 (2018): 920–980.

Clemens, Michael A. "The meaning of failed replications: A review and proposal." Journal of Economic Surveys 31, no. 1 (2017): 326–342.

Coile, Courtney C., and Phillip B. Levine. "Recessions, Retirement, and Social Security." American Economic Review 101, no. 3 (2011): 23–28.

Comolli, Chiara Ludovica, and Fabrizio Bernardi. "The causal effect of the great recession on childlessness of white American women." IZA Journal of Labor Economics 4, no. 1 (2015): 1–21.

Condliffe, Simon, Matt B. Saboe, and Sabrina Terrizzi. "Did the ACA reduce job-lock and spur entrepreneurship?" Journal of Entrepreneurship and Public Policy 6, no. 2 (2017): 150–163.

Costa, Dora L., and Matthew E. Kahn. Why has California's residential electricity consumption been so flat since the 1980s? A microeconometric approach. No. w15978. National Bureau of Economic Research, 2010.

Costa, Dora L., and Matthew E. Kahn. "Electricity consumption and durable housing: Understanding cohort effects." American Economic Review: Papers & Proceedings 101, no. 3 (2011): 88–92.

Couture, Victor, and Jessie Handbury. "Urban revival in America." Journal of Urban Economics 119 (September 2020): 103267.

Culhane, Dennis P., Stephen Metraux, and Trevor Hadley. "Public service reductions associated with placement of homeless persons with severe mental illness in supportive housing." Housing Policy Debate 13, no. 1 (2002): 107–163.

Dahl, Gordon B., and Enrico Moretti. "The demand for sons." The Review of Economic Studies 75, no. 4 (2008): 1085–1120.

Davila, Alberto, and Marie T. Mora. "Changes in the earnings of Arab men in the US between 2000 and 2002." Journal of Population Economics 18, no. 4 (2005): 587–601.

Dillender, Marcus. "Do more health insurance options lead to higher wages? Evidence from states extending dependent coverage?" Journal of Health Economics 36 (July, 2014): 84–97.

Dunning, Thad. Natural experiments in the social sciences: A design-based approach. Cambridge University Press, 2012.

Elliott, Catherine S. "A May American Economic Review papers seminar and an analytic project for advanced undergraduates." Journal of Economic Education 35, no. 3 (2004): 232.

Fowlie, Meredith, Michael Greenstone, and Catherine Wolfram. "Do energy efficiency investments deliver? Evidence from the weatherization assistance program." The Quarterly Journal of Economics 133, no. 3 (2018): 1597–1644.

Frean, Molly, Jonathan Gruber, and Benjamin D. Sommers. "Premium subsidies, the mandate, and medicaid expansion: Coverage effects of the Affordable Care Act." Journal of Health Economics 53 (2017): 72–86.

Frey, William H. "Will this be the decade of big city growth?" May 23, 2014. https://www.brookings.edu/opinions/will-this-be-the-decade-of-bigcity-growth/.

Frey, William H. "American cities saw uneven growth last decade, new census data show." May 26, 2020. https://www.brookings.edu/research/newcensus-data-show-an-uneven-decade-of-growth-for-us-cities/.

Gerring, John. "Mere description." British Journal of Political Science (2012): 721–746.

Ginther, D. K., and M. Zavodny. "Is the male marriage premium due to selection? The effect of shotgun weddings on the return to marriage." Journal of Population Economics 14, no. 2 (2001): 313–328.

Gillingham K. Rebound Effects. In: Palgrave Macmillan (eds.), The New Palgrave Dictionary of Economics. London: Palgrave Macmillan, 2014. https://doi.org/10.1057/978-1-349-95121-5_2875-1.

Goodman-Bacon, Andrew. "Difference-in-differences with variation in treatment timing." Working Paper. July, 2019.

Grazi, F., J. C. van den Bergh, and J. N. van Ommeren. "An empirical analysis of urban form, transport, and global warming." The Energy Journal 29, no. 4 (2008): 97–123.

Grimmer, Justin. "We are all social scientists now: How big data, machine learning, and causal inference work together." PS, Political Science & Politics 48, no. 1 (2015): 80.

Hanushek, Eric A. "The failure of input-based schooling policies." The Economic Journal 113, no. 485 (2003): F64–F98.

Holian, Matthew. "Where is the city's center? Five measures of central location." Cityscape: A Journal of Policy Development and Research 21, no. 2 (2019).

Holian, Matthew J. "The impact of urban form on vehicle ownership." Economics Letters 186 (2020a): 108763.

Holian, Matthew J. "The impact of building energy codes on household electricity expenditures." Economics Letters 186 (2020b): 108841.

Holian, Matthew J. "Corrigendum to 'The impact of building energy codes on household electricity expenditures' [Econom. Lett. 186 (2020) 108841]." Economics Letters 200 (2021): 109738.

Huang, Eddie. Fresh off the boat: A memoir. Spiegel & Grau, 2013.

Jacobsen, Grant D., and Matthew J. Kotchen. "Are building codes effective at saving energy? Evidence from residential billing data in Florida." Review of Economics and Statistics 95, no. 1 (2013): 34–49.

Koirala, Bishwa S., Alok K. Bohara, and Hui Li. "Effects of energy-efficiency building codes in the energy savings and emissions of carbon dioxide." Environmental Economics and Policy Studies 15, no. 3 (2013): 271–290.

Kostandini, Genti, Elton Mykerezi, and Cesar Escalante. "The impact of immigration enforcement on the US farming sector." American Journal of Agricultural Economics 96, no.1 (2013): 172–192.

Kotchen, Matthew J. "Longer-run evidence on whether building energy codes reduce residential energy consumption." Journal of the Association of Environmental and Resource Economists 4, no. 1 (2017): 135–153.

Krueger, Alan B. "Economic considerations and class size." The Economic Journal 113, no. 485 (2003): F34–F63.

Kuka, Elira, Na'ama Shenhav, and Kevin Shih. "A reason to wait: The effect of legal status on teen pregnancy." AEA Papers and Proceedings 109 (2019): 213–217.

Lee, Jun Yeong, and John V. Winters. "State medicaid expansion and the selfemployed." IZA Discussion Paper No. 12997 (2020).

Levinson, Arik. "How much energy do building energy codes save? Evidence from California houses." American Economic Review 106, no. 10 (2016): 2867–2894.

Manning, Matthew, Shane D. Johnson, Nick Tilley, Gabriel T. W. Wong, and Margarita Vorsina. Economic analysis and efficiency in policing, criminal justice and crime reduction: What works? London: Palgrave Macmillan, 2016.

Marshall, M. I., and A. Flaig. "Marriage, children, and self-employment earnings: An analysis of self-employed women in the US." Journal of Family and Economic Issues 35, no. 3 (2014): 313–322.

Munger, M. C. Tomorrow 3.0: Transaction costs and the sharing economy. Cambridge University Press, 2018.

Novan, Kevin, Aaron Smith, and Tianxia Zhou. "Residential building codes do save energy: Evidence from hourly smart-meter data." UC Davis (2017).

Nordhaus, William D. The climate casino: Risk, uncertainty, and economics for a warming world. Yale University Press, 2013.

Nordhaus, William D. Revisiting the social cost of carbon PNAS 114, no. 7 (February 14, 2017): 1518–1523. https://www.pnas.org/content/114/7/1518.

Orrenius, Pia M., and Madeline Zavodny. Beside the golden door: US immigration reform in a new era of globalization. AEI Press, 2010.

Orrenius, Pia M., and Madeline Zavodny. "The impact of temporary protected status on immigrants' labor market outcomes." American Economic Review: Papers & Proceedings 105, no. 5 (2015): 576–580.

Rossin-Slater, Maya, Christopher J. Ruhm, and Jane Waldfogel. "The effects of California's paid family leave program on mothers? Leave? Taking and subsequent labor market outcomes." Journal of Policy Analysis and Management 32, no. 2 (2013): 224–245.

Ruggles, Steven. "Big microdata for population research." Demography 51, no. 1 (2014): 287–297.

Ruggles, Steven, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas, and Matthew Sobek. IPUMS USA: Version 10.0 [dataset]. Minneapolis, MN: IPUMS, 2020. https://doi.org/10.18128/D010.V10.0.

Sastry, N., and Gregory, J. "The location of displaced New Orleans residents in the year after Hurricane Katrina." Demography 51, no. 3 (2014): 753–775.

Schneebaum, Alyssa, and M.V. Lee Badgett. "Poverty in US lesbian and gay couple households." Feminist Economics 25, no. 1 (2019): 1–30.

Sjoquist, D. L., and J. V. Winters. "State merit aid programs and college major: A focus on STEM." Journal of Labor Economics 33, no. 4 (2015a): 973–1006.

Sjoquist, D. L., and J. V. Winters. "State merit-based financial aid programs and college attainment." Journal of Regional Science 55, no. 3 (2015b): 364–390.

Stansel, Dean, Gary Jackson, and Howard Finch. "Housing tenure and mobility with an acquisition-based property tax: The case of Florida." Journal of Housing Research 16, no. 2 (2007): 117–129.

Stock, James H., and Mark W. Watson. Introduction to econometrics. Boston, MA: Addison Wesley, 2011.

Sturtevant, Lisa. "The new District of Columbia: What population growth and demographic change mean for the city." Journal of Urban Affairs 36, no. 2 (2014): 276–299.

Thornton, Robert J., and Edward J. Timmons. "Licensing one of the world's oldest professions: Massage." The Journal of Law and Economics 56, no. 2 (2013): 371–388.

Walker, K. E. "Baby boomer migration and demographic change in US metropolitan areas." Migration Studies 4, no. 3 (2016): 347–372.

Winters, John V. "Do earnings by college major affect graduate migration?" The Annals of Regional Science 59, no. 3 (2017): 629–649.
Winters, J. V. "Is economics a good major for future lawyers? Evidence from earnings data." The Journal of Economic Education 47, no. 2 (2016): 187–191.

Author Index

M. J. Holian, *Data and the American Dream*,
https://doi.org/10.1007/978-3-030-64262-4

Subject Index

M. J. Holian, *Data and the American Dream*,
https://doi.org/10.1007/978-3-030-64262-4