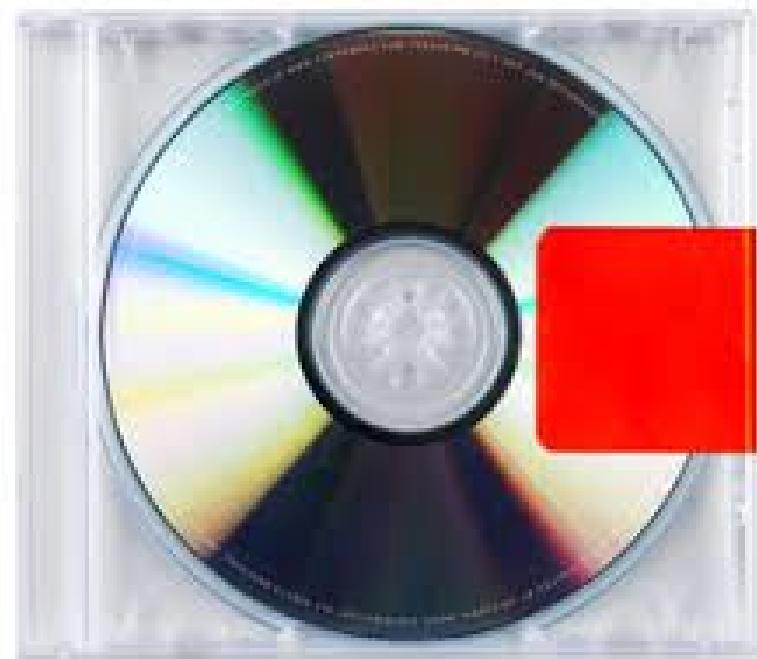


SCOTT CUNNINGHAM

CAUSAL INFERENCE: THE MIXTAPE (V. 1.8)



Copyright © 2020 Scott Cunningham

PUBLISHED BY

TUFTE-LATEX.GOOGLECODE.COM

Licensed under the Apache License, Version 2.0 (the “License”); you may not use this file except in compliance with the License. You may obtain a copy of the License at <http://www.apache.org/licenses/LICENSE-2.0>. Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an “AS IS” BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the License for the specific language governing permissions and limitations under the License.

First printing, March 2020

Contents

<i>Introduction</i>	13
<i>Probability theory and statistics review</i>	23
<i>Properties of Regression</i>	35
<i>Directed acyclical graphs</i>	67
<i>Potential outcomes causal model</i>	81
<i>Matching and subclassification</i>	105
<i>Regression discontinuity</i>	151
<i>Instrumental variables</i>	203
<i>Panel data</i>	241
<i>Differences-in-differences</i>	259
<i>Synthetic control</i>	283

Conclusion 309

Bibliography 311

List of Figures

- 1 xkcd 16
- 2 Wright's graphical demonstration of the identification problem 20
- 3 Graphical representation of bivariate regression from y on x 46
- 4 Distribution of residuals around regression line 47
- 5 Distribution of coefficients from Monte Carlo simulation. 54
- 6 Regression anatomy display. 61
- 7 Top left figure: Non-star sample scatter plot of beauty (vertical axis) and talent (horizontal axis). Top right figure: Star sample scatter plot of beauty and talent. Bottom left figure: Entire (stars and non-stars combined) sample scatter plot of beauty and talent. 79
- 8 Regression of kindergarten percentile scores onto treatments [Krueger, 1999]. 98
- 9 Regression of first grade percentile scores onto treatments [Krueger, 1999]. 99
- 10 Regression of first grade percentile scores onto treatments for K-3 with imputed test scores for all post-kindergarten ages [Krueger, 1999]. 100
- 11 Switching of students into and out of the treatment arms between first and second grade [Krueger, 1999]. 101
- 12 IV reduced form approach compared to the OLS approach [Krueger, 1999]. 102
- 13 Lung cancer at autopsy trends 107
- 14 Smoking and Lung Cancer 109
- 15 Covariate distribution by job trainings and control. 122
- 16 Covariate distribution by job trainings and matched sample. 125
- 17 Lalonde [1986] Table 5(a) 135
- 18 Lalonde [1986] Table 5(b) 136
- 19 Dehejia and Wahba [1999] Figure 1, overlap in the propensity scores (using PSID) 137
- 20 Dehejia and Wahba [1999] Figure 2, overlap in the propensity scores (using CPS) 138

- 21 Dehejia and Wahba [1999] Table 3 results. 138
 22 Dehejia and Wahba [1999] Table 4, covariate balance 139
 23 Histogram of propensity score by treatment status 142
- 24 Angrist and Lavy [1999] descriptive statistics 155
 25 Angrist and Lavy [1999] descriptive statistics for the discontinuity sample. 155
 26 Maimonides' Rule vs. actual class size [Angrist and Lavy, 1999]. 156
 27 Average reading scores vs. enrollment size [Angrist and Lavy, 1999]. 157
 28 Reading score residual and class size function by enrollment count [Angrist and Lavy, 1999]. 158
 29 Math score residual and class size function by enrollment count [Angrist and Lavy, 1999]. 158
 30 OLS regressions [Angrist and Lavy, 1999]. 159
 31 First stage regression [Angrist and Lavy, 1999]. 160
 32 Second stage regressions [Angrist and Lavy, 1999]. 160
 33 Sharp vs. Fuzzy RDD [van der Klaauw, 2002]. 162
 34 Dashed lines are extrapolations 163
 35 Display of observations from simulation. 165
 36 Display of observations discontinuity simulation. 166
 37 Simulated nonlinear data from Stata 167
 38 Illustration of a boundary problem 169
 39 Insurance status and age 170
 40 Card et al. [2008] Table 1 173
 41 Investigating the CPS for discontinuities at age 65 [Card et al., 2008] 174
 42 Investigating the NHIS for the impact of Medicare on care and utilization [Card et al., 2008] 175
 43 Changes in hospitalizations [Card et al., 2008] 175
 44 Mortality and Medicare [Card et al., 2009] 176
 45 Imbens and Lemieux [2008], Figure 3. Horizontal axis is the running variable. Vertical axis is the conditional probability of treatment at each value of the running variable. 177
 46 Potential and observed outcome regressions [Imbens and Lemieux, 2008] 177
 47 Panel C is density of income when there is no pre-announcement and no manipulation. Panel D is the density of income when there is pre-announcement and manipulation. From McCrary [2008]. 180
 48 Panels refer to (top left to bottom right) district characteristics: real income, percent high school degree, percent black, and percent eligible to vote. Circles represent the average characteristic within intervals of 0.01 in Democratic vote share. The continuous line represents the predicted values from a fourth-order polynomial in vote share fitted separately for points above and below the 50 percent threshold. The dotted line represents the 95 percent confidence interval. 182
 49 Example of outcome plotted against the running variable. 182

- 50 Example of covariate plotted against the running variable. 183
 51 McCrary density test, NHIS data, SNAP eligibility against a running variable based on income and family size. 184
 52 Lee, Moretti and Butler (2004)'s Table 1. Main results. 187
 53 Lee et al. [2004], Figure I 192
 54 Reproduction of Lee et al. [2004] Figure I using `cmogram` with quadratic fit and confidence intervals 193
 55 Reproduction of Lee et al. [2004] Figure I using `cmogram` with linear fit 194
 56 Reproduction of Lee et al. [2004] Figure I using `cmogram` with lowess fit 195
 57 Carrell et al. [2011] Figure 3 196
 58 Carrell et al. [2011] Table 3 196
 59 Local linear nonparametric regressions 197
 60 Local linear nonparametric regressions 199
 61 RKD kinks from Card et al. [2015] 200
 62 Base year earnings and benefits for single individuals from Card et al. [2015] 201
 63 Log(duration unemployed) and benefits for single individuals from Card et al. [2015] 201
 64 Pseudoephedrine (top) vs d-methamphetamine (bottom) 218
 65 Figure 3 from Cunningham and Finlay [2012] showing changing street prices following both supply shocks. 219
 66 Figure 5 from Cunningham and Finlay [2012] showing first stage. 220
 67 Figure 4 from Cunningham and Finlay [2012] showing reduced form effect of interventions on children removed from families and placed into foster care. 220
 68 Table 3 Cunningham and Finlay [2012] showing OLS and 2SLS estimates of meth on foster care admissions. 221
 69 Angrist and Krueger [1991] explanation of their instrumental variable. 223
 70 Angrist and Krueger [1991] first stage relationship between quarter of birth and schooling. 224
 71 Angrist and Krueger [1991] reduced form visualization of the relationship between quarter of birth and log weekly earnings. 224
 72 Angrist and Krueger [1991] first stage for different outcomes. 225
 73 Angrist and Krueger [1991] OLS and 2SLS results for the effect of education on log weekly earnings. 227
 74 Bound et al. [1995] OLS and 2SLS results for the effect of education on log weekly earnings. 229
 75 Bound et al. [1995] OLS and 2SLS results for the effect of education on log weekly earnings with the 100+ weak instruments. 230

76	Table 3 from Cornwell and Trumbull [1994]	248
77	Table 2 from Draca et al. [2011]	249
78	Table 2 from Cornwell and Rupert [1997]	250
79	Table 3 from Cornwell and Rupert [1997]	251
80	NJ and PA	262
81	Distribution of wages for NJ and PA in November 1992	263
82	Simple DD using sample averages	264
83	DD regression diagram	266
84	Checking the pre-treatment trends for parallelism	267
85	Autor [2003] leads and lags in dynamic DD model	267
86	Gruber [1994] Table 3	271
87	Internet diffusion and music spending	272
88	Comparison of Internet user and non-user groups	273
89	Theoretical predictions of abortion legalization on age profiles of gonorrhea incidence	275
90	Differences in black female gonorrhea incidence between repeal and Roe cohorts.	276
91	Coefficients and standard errors from DD regression equation	277
92	Subset of coefficients (year-repeal interactions) for the DDD model, Table 3 of Cunningham and Cornwell [2013].	278
93	I ❤️ Federalism bumpersticker (for the natural experiments)	280
94	California cigarette sales vs the rest of the country	288
95	California cigarette sales vs synthetic California	288
96	Balance table	289
97	California cigarette sales vs synthetic California	290
98	Placebo distribution	291
99	Placebo distribution	292
100	Placebo distribution	292
101	West Germany GDP vs. Other Countries	293
102	Synthetic control graph: West Germany vs Synthetic West Germany	294
103	Synthetic control graph: Differences between West Germany and Synthetic West Germany	294
104	Synthetic control graph: Placebo Date	295
105	Prison capacity (operational capacity) expansion	296
106	African-American male incarceration rates	297
107	African-American male incarceration	299
108	Gap between actual Texas and synthetic Texas	300
109	Histogram of the distribution of ratios of post-RMSPE to pre-RMSPE. Texas is one of the ones in the far right tail.	305
110	Histogram of the distribution of ratios of post-RMSPE to pre-RMSPE. Texas is one of the ones in the far right tail.	306
111	Placebo distribution. Texas is the black line.	308

List of Tables

1	Examples of Discrete and Continuous Random Processes.	23
2	Total number of ways to get a 7 with two six-sided dice.	25
3	Total number of ways to get a 3 using two six-sided dice.	25
4	Two way contingency table.	28
5	Sum of deviations equalling zero	37
6	Simulated data showing the sum of residuals equals zero	48
7	Monte Carlo simulation of OLS	54
8	Regressions illustrating confounding bias with simulated gender disparity	77
9	Regressions illustrating collider bias	77
10	Yule regressions [Yule, 1899].	82
11	Potential outcomes for ten patients receiving surgery Y^1 or chemo Y^0 .	88
12	Post-treatment observed lifespans in years for surgery $D = 1$ versus chemotherapy $D = 0$.	89
13	Krueger regressions [Krueger, 1999].	103
14	Death rates per 1,000 person-years [Cochran, 1968]	110
15	Mean ages, years [Cochran, 1968].	111
16	Subclassification example.	112
17	Adjusted mortality rates using 3 age groups [Cochran, 1968].	112
18	Subclassification example of Titanic survival for large K	117
19	Training example with exact matching	120
20	Training example with exact matching (including matched sample)	123
21	Another matching example (this time to illustrate bias correction)	129
22	Nearest neighbor matched sample	130
23	Nearest neighbor matched sample with fitted values for bias correction	131
24	Completed matching example with single covariate	137
25	Distribution of propensity score for treatment group.	141
26	Distribution of propensity score for CPS Control group.	141

27	Balance in covariates after coarsened exact matching.	149
28	OLS and 2SLS regressions of Log Earnings on Schooling	236
29	OLS and 2SLS regressions of Log Quantity on Log Price with wave height instrument	238
30	OLS and 2SLS regressions of Log Quantity on Log Price with wind-speed instrument	239
31	POLS, FE and Demeaned OLS Estimates of the Determinants of Log Hourly Price for a Panel of Sex Workers	255
32	Compared to what? Different cities	260
33	Compared to what? Before and After	260
34	Compared to what? Subtract each city's differences	260
35	Differences-in-differences-in-differences	269
36	Synthetic control weights	301

*To my son, Miles, one of my favorite people. I
love you. You've tagged my head and heart.*

Introduction

I like to think of causal inference as the space between theory and estimation. It's where we test primarily social scientific hypotheses *in the wild*. Some date the beginning of modern causal inference with Fisher [1935], Haavelmo [1943], Rubin [1974] or applied labor economics studies; but whenever you consider its start, causal inference is now a distinct field within econometrics. It's sometimes listed as a lengthy chapter on "program evaluation" [Wooldridge, 2010], or given entire book-length treatments. To name just a few textbooks in the growing area, there's Angrist and Pischke [2009], Morgan and Winship [2014], Imbens and Rubin [2015] and probably a half dozen others, not to mention numerous, lengthy treatments of specific strategies such as Imbens and Lemieux [2008] and Angrist and Krueger [2001]. The field is crowded and getting more crowded every year.

So why does my book exist? I believe there's some holes in the market, and this book is an attempt to fill them. For one, none of the materials out there at present are exactly what I need when I teach my own class on causal inference. When I teach that class, I use Morgan and Winship [2014], Angrist and Pischke [2009], and a bunch of other stuff I've cobbled together. No single book at present has everything I need or am looking for. Imbens and Rubin [2015] covers the potential outcomes model, experimental design, matching and instrumental variables, but does not contain anything about directed acyclic graphical models, regression discontinuity, panel data or synthetic control.¹ Morgan and Winship [2014] covers DAGs, the potential outcomes model, and instrumental variables, but is missing adequate material on regression discontinuity, panel data and synthetic control. Angrist and Pischke [2009] is very close, but does not include anything on synthetic control nor the graphical models that I find so useful. But maybe most importantly, Imbens and Rubin [2015], Angrist and Pischke [2009] and Morgan and Winship [2014] do not provide enough practical guidance in Stata, which I believe is invaluable for learning and becoming confident in this area.²

This book was written for a few different people. It was written

¹ But hopefully volume 2 will build on volume 1 and continue to build out this material, at which point my book becomes obsolete.

² Although Angrist and Pischke [2009] provides an online data warehouse from dozens of papers, I find that students need more pedagogical walkthroughs and replications for these ideas to become concrete and familiar.

first and foremost for *practitioners*, which is why it includes easy to download datasets and programs. It's why I have made several efforts to review papers as well as replicate them as much as possible. I want readers to both understand this field, but also importantly, to feel empowered to apply these methods and techniques to their own research questions.

Another person I have in mind is the experienced social scientist wanting to retool. Maybe these are people with more of a theoretical bent or background, or maybe it's people who simply have some holes in their human capital. This book, I hope, can help guide them through the modern theories of causality so common in the social sciences, as well as provide a calculus in directed acyclic graphical models that can help connect their knowledge of theory with econometric identification.

Finally, this book is written for people very early in their careers, be it undergraduates, graduate students, or newly minted PhDs. My hope is that this book can give you a jump start so that you don't have to, like many of us had to, meander through a somewhat labyrinthine path to these methods.

Giving it away

For now, I have chosen to give this book away, for several reasons. First, the most personal ones. I derive satisfaction in knowing that I can take what I've learned, and my personal philosophies about this material, including how it's taught, and give it away to people. This is probably because I remain deep down a teacher who cares about education. I love helping students discover; I love sharing in that discovery. And if someone is traveling the same windy path that I traveled, then why not help them by sharing what I've learned and now believe about this field? I could sell it, and maybe one day I will, but for the moment I've decided to give it away – at least, the first few versions.

The second reason, which supports the first, is something that Al Roth once told me. He had done me a favor, which I could never repay, and I told him that. To which he said:

“Scott, intergenerational favors aren’t supposed to be repaid, they’re supposed to be passed forward to the next generation.”

I've given a lot of thought to what he said³, and if you'll indulge me, I'd like to share my response to what Roth said to me. Every person must decide what their values are, how they want to live their life, and what they want to say about the life they were given to live

³ I give a lot of thought to anything and everything that Roth says or has ever said actually.

when they look back on it. Economic models take preferences as given and unchanging [Becker, 1993], but I have found that figuring out one's preferences is the hard work of being a moral person.

Love for others, love for my students, love for my senior mentors, love for my friends, love for my peers, love for junior faculty, love for graduate students, love for my family – these are the things that motivate me. I want my life to be of service to others, but I'm a teacher and a researcher, not Mother Theresa. So I have to figure out what it means to be of service to others as a teacher and a researcher, given that is a major part of my life. Each of us have to figure out what it means to be a neighbor with the resources we've been given and the path we've chosen. So, somewhat inspired by Roth and various senior mentors' generosities towards me, I decided that at least for now giving away the book is one very small way to live consistently with these values.

Plus, and maybe this is actually one of the more important reasons, I figure if I give away the book, then you, the reader, will be patient with me as I take my time to improve the book. Not everything is in this book. I see it as foundational, not comprehensive. A useful starting point, not an ending point. If you master the material in this book, then you'll have a solid foundation to build on. You might explore the exciting new area of causal inference and machine learning by Athey, Imbens and others, structural econometrics [Rust, 1987], synthetic control with multiple treatments [Cavallo et al., 2013], randomized controlled trials and field experiments, and the seemingly never-ending innovations and updates in econometrics.

Another more playful reason I am giving it away is because I find Chance the Rapper's mentality when it comes to mixtapes infectious. A mixtape is a collection of one's favorite songs given away to friends in the hopes they'll have fun listening to it. Consider this my mixtape of research designs that I hope you'll find fun, interesting, and powerful for estimating causal effects. It's not everything you need to know; more like the seminal things you should know as of this book's writing. There's far more to learn, and I consider this to be the first book you need, not the only book you need. This book is meant to be a complement to books like Angrist and Pischke [2009] rather than a substitute.

How I got here

It may be interesting to hear how I got to the point of wanting to write this book. The TL;DR version is that I followed a windy path from poetry to economics to research. I fell in love with economics,

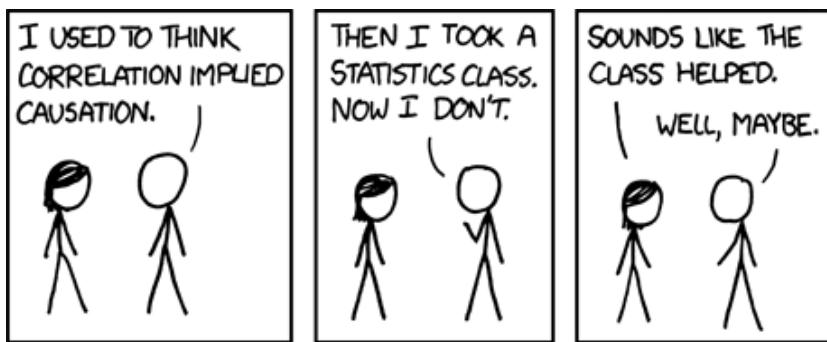


Figure 1: xkcd

then research, and causal inference was a constant throughout all of it. But now the longer version.

I majored in English at the University of Tennessee at Knoxville and graduated with a serious ambition of becoming a professional poet. But, while I had been successful writing poetry in college, I quickly realized that the road to success beyond that point was probably not realistic. I was newly married with a baby on the way, and working as a qualitative research analyst doing market research and slowly stopped writing poetry altogether.⁴

My job as a qualitative research analyst was eye opening in part because it was my first exposure to empiricism. My job was to do “grounded theory” – a kind of inductive approach to generating explanations of human behavior based on focus groups and in-depth interviews, as well as other ethnographic methods. I approached each project as an opportunity to understand why people did the things they did (even if what they did was buy detergent or pick a cable provider). While the job inspired me to develop my own theories about human behavior, it didn’t provide me a way of falsifying those theories.

I lacked a background in the social sciences, so I would spend my evenings downloading and reading articles from the Internet. I don’t remember how I ended up there, but one night I was on the University of Chicago Law and Economics working paper series when a speech by Gary Becker caught my eye. It was his Nobel Prize acceptance speech on how economics applied to all of human behavior [Becker, 1993], and reading it changed my life. I thought economics was about stock markets and banks until I read that speech. I didn’t know economics was an engine that one could use to analyze all of human behavior. This was overwhelmingly exciting, and a seed had been planted.

But it wasn’t until I read Lott and Mustard [1997] that I became

⁴ Rilke said you should quit writing poetry when you can imagine yourself living without it [Rilke, 2012]. I could imagine living without poetry, so I took his advice and quit. I have no regrets whatsoever. Interestingly, when I later found economics, I believed I would never be happy unless I was a professional economist doing research on the topics I found interesting. So I like to think I followed Rilke’s advice on multiple levels.

truly enamored with economics. I had no idea that there was an empirical component where economists sought to estimate causal effects with quantitative data. One of the authors in [Lott and Mustard \[1997\]](#) was David Mustard, then an Associate Professor of economics at the University of Georgia, and one of Becker's former students. I decided that I wanted to study with Mustard, and therefore applied for University of Georgia's doctoral program in economics. I moved to Athens, Georgia with my wife, Paige, and our infant son, Miles, and started classes in the fall of 2002.

After passing my prelims, I took Mustard's labor economics field class, and learned about the kinds of topics that occupied the lives of labor economists. These topics included the returns to education, inequality, racial discrimination, crime and many other fascinating and important topics. We read many, many empirical papers in that class, and afterwards I knew that I would need a strong background in econometrics to do the kind of empirical work I desired to do. And since econometrics was the most important area I could ever learn, I decided to make it my main field of study. This led to me working with Christopher Cornwell, an econometrician at Georgia from whom I learned a lot. He became my most important mentor, as well as a coauthor and friend. Without him, I wouldn't be where I am today.

Econometrics was difficult. I won't pretend I was a prodigy. I took all the econometrics courses offered at the University of Georgia, and some more than once. They included probability and statistics, cross-section, panel data, time series, and qualitative dependent variables. But while I passed my field exam in econometrics, I failed to understand econometrics at deeper, more basic levels. You might say I lost sight of the forest for the trees.

I noticed something while I was writing the third chapter of my dissertation that I hadn't noticed before. My third chapter was an investigation of the effect of abortion legalization on longrun risky sexual behavior [[Cunningham and Cornwell, 2013](#)]. It was a revisiting of [Donohue and Levitt \[2001\]](#). One of the books I read in preparation of the study was [Levine \[2004\]](#), which in addition to reviewing the theory of and empirical studies on abortion had a little table explaining the differences-in-differences identification strategy. The University of Georgia had a traditional econometrics pedagogy, and most of my field courses were theoretical (e.g., Public and Industrial Organization), so I never really had heard the phrase "identification strategy" let alone "causal inference". That simple difference-in-differences table was eye-opening. I saw how econometric modeling could be used to isolate the causal effects of some treatment, and that put me on a new research trajectory.

Optimization Makes Everything Endogenous

Causal inference is often accused of being a-theoretical, but nothing could be further from the truth [Imbens, 2009, Deaton and Cartwright, 2018]. Economic theory is *required* in order to justify a credible claim of causal inference. And economic theory also highlights why causal inference is necessarily a thorny task. Let me explain.

There's broadly thought to be two types of data. There's experimental data and non-experimental data. The latter is also sometimes called *observational* data. Experimental data is collected in something akin to a laboratory environment. In a traditional experiment, the researcher participates actively in the process being recorded. It's more difficult to obtain data like this in the social sciences due to feasibility, financial cost or moral objections, although it is more common now than was once the case. Examples include the Oregon Medicaid Experiment, the RAND health insurance experiment, the field experiment movement inspired by Michael Kremer, Esther Duflo and John List, and many others.

Observational data is usually collected through surveys on a retrospective manner, or as the byproduct of some other business activity ("big data"). That is, in observational studies, you collect data about what has happened previously, as opposed to collecting data as it happens. The researcher is also a passive actor in this process. She observes actions and results, but is not in a position to interfere with the outcomes. This is the most common form of data that social scientists use.

Economic theory tells us we should be suspicious of correlations found in observational data. In observational data, correlations are almost certainly not reflecting a causal relationship because the variables were endogenously chosen by people who were making decisions they thought were best. In pursuing some goal while facing constraints, they chose certain things that created a spurious correlation with other things. The reason we think is because of what we learn from the potential outcomes model: a correlation, in order to be a measure of a causal effect, must be completely independent of the potential outcomes under consideration. Yet if the person is making some choice *based* on what she thinks is best, then it necessarily violates this independence condition. Economic theory predicts choices will be endogenous, and thus naive correlations are misleading.

But theory, combined with intimate knowledge of the institutional details surrounding the phenomena under consideration, can be used to recover causal effects. We can estimate causal effects, but only with

assumptions and data.

Now we are veering into the realm of epistemology. Identifying causal effects involves assumptions, but it also requires a particular kind of belief about the work of scientists. Credible and valuable research requires that we believe that it is more important to do our work *correctly* than to try and achieve a certain outcome (e.g., confirmation bias, statistical significance, stars). The foundations of scientific knowledge are scientific methodologies. Science does not collect evidence in order to prove what we want to be true or what people want others to believe. That is a form of *propaganda*, not science. Rather, scientific methodologies are devices for forming a particular kind of belief. Scientific methodologies allow us to accept unexpected, and sometimes, undesirable answers. They are process oriented, not outcome oriented. And without these values, causal methodologies are also not credible.

Example: Identifying price elasticity of demand

One of the cornerstones of scientific methodologies is empirical analysis.⁵ By empirical analysis, I mean the use of data to test a theory or to estimate a relationship between variables. The first step in conducting an empirical economic analysis is the careful formulation of the question we would like to answer. In some cases, we would like to develop and test a formal economic model that describes mathematically a certain relationship, behavior or process of interest. Those models are valuable insofar as they both describe the phenomena of interest as well as make falsifiable (testable) predictions. A prediction is falsifiable insofar as we can evaluate, and potentially reject the prediction, with data.⁶ The economic model is the framework with which we describe the relationships we are interested in, the intuition for our results and the hypotheses we would like to test.⁷

After we have specified an economic model, we turn it into what is called an econometric model that we can estimate directly with data. One clear issue we immediately face is regarding the functional form of the model, or how to describe the relationships of the variables we are interested in through an equation. Another important issue is how we will deal with variables that cannot be directly or reasonably observed by the researcher, or that cannot be measured very well, but which play an important role in our economic model.

A generically important contribution to our understanding of causal inference is the notion of comparative statics. Comparative statics are theoretical descriptions of causal effects contained within the model. These kinds of comparative statics are always based on

⁵ It is not the only cornerstone, nor even necessarily the most important cornerstone, but empirical analysis has always played an important role in scientific work.

⁶ You can also obtain a starting point for empirical analysis less formally through an intuitive and less formal reasoning process. But economics favors formalism and deductive methods.

⁷ Economic models are abstract, not realistic, representations of the world. George Box, the statistician, once quipped that “all models are wrong, but some are useful.”

the idea of *ceteris paribus* – holding all else constant. When we are trying to describe the causal effect of some intervention, we are always assuming that the other relevant variables in the model are not changing. If they weren't, then it confounds our estimation.

One of the things implied by *ceteris paribus* that comes up repeatedly in this book is the idea of covariate balance. If we say that everything is the same except for the movement of one variable, then everything is the same on both sides of that variable's changing value. Thus, when we invoke *ceteris paribus*, we are implicitly invoking covariate balance – both the observable and unobservable covariates.

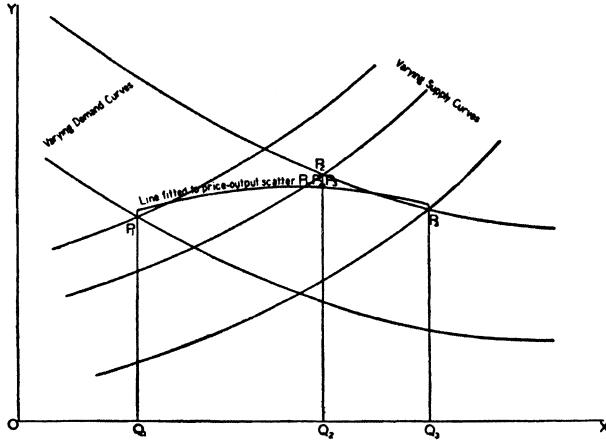
To illustrate this idea, let's begin with a basic economic model: supply and demand equilibrium and the problems it creates for estimating the price elasticity of demand. Policy-makers and business managers have a natural interest in learning the price elasticity of demand. Knowing this can help firms maximize profits, help governments choose optimal taxes, as well as the conditions under which quantity restrictions are preferred [Becker et al., 2006]. But, the problem is that we do not observe demand curves, because demand curves are theoretical objects. More specifically, a demand curve is a collection of paired potential outcomes of price and quantity. We observe *price and quantity equilibrium values*, not the potential price and potential quantities along the entire demand curve. Only by tracing out the potential outcomes along a demand curve can we calculate the elasticity.

To see this, consider this graphic from Philip Wright's Appendix B [Wright, 1928], which we'll discuss in greater detail later (Figure 2). The price elasticity of demand is the ratio of percentage changes in quantity to price *for a single demand curve*. Yet, when there are shifts in supply and demand, a sequence of quantity and price pairs emerge in history which reflect neither the demand curve nor the supply curve. In fact, connecting the points does not reflect any meaningful object.

The price elasticity of demand is the solution to the following equation:

$$\epsilon = \frac{\partial \log Q}{\partial \log P}$$

But in this example, the change in P is *exogenous*. For instance, it holds supply fixed, the prices of other goods fixed, income fixed, preferences fixed, input costs fixed, etc. In order to estimate the price elasticity of demand, we need changes in P that are completely and utterly independent of the otherwise normal determinants of supply and the other determinants of demand. Otherwise we get shifts in either supply or demand, which creates new pairs of data for which any correlation between P and Q will not be a measure of the

*Exhibit 1***The Graphical Demonstration of the Identification Problem in Appendix B (p. 296)****FIGURE 4. PRICE-OUTPUT DATA FAIL TO REVEAL EITHER SUPPLY OR DEMAND CURVE.****Figure 2: Wright's graphical demonstration of the identification problem**

elasticity of demand.

Nevertheless, the elasticity is an important object, and we need to know it. So given this theoretical object, we must write out an econometric model as a starting point. One possible example of an econometric model would be a linear demand function:

$$\log Q_d = \alpha + \delta \log P + \gamma X + u$$

where α is the intercept, δ is the elasticity of demand, X is a matrix of factors that determine demand like the prices of other goods or income γ is the coefficient on the relationship between X and Q_d and u is the error term.⁸

Foreshadowing the rest of the lecture notes, to estimate price elasticity of demand, we need two things. First, we need numerous rows of data on price and quantity. Second, we need for the variation in price in our imaginary dataset to be independent of u . We call this kind of independence *exogeneity*. Without both, we cannot recover the price elasticity of demand, and therefore any decision that requires that information will be based on flawed or incomplete data.

In conclusion, simply finding an association between two variables might be suggestive of a causal effect, but it also might not. Correlation doesn't mean causation unless key assumptions hold. Before we start digging into causal inference, we need to lay down a foundation in simple regression models. We also need to introduce a simple program necessary to do some of the Stata examples. I have uploaded numerous datasets to my website which you will use to

⁸ More on the error term later.

perform procedures and replicate examples from the book. That file can be downloaded from <http://scunning.com/scuse.ado>. Once it's downloaded, simply copy and paste the file into your personal Stata ado subdirectory.⁹ You're all set to go forward!

⁹ To find that path, type in `sysdir` at the Stata command line. This will tell you the location of your personal ado subdirectory. If you copy `scuse.ado` it into this subdirectory, then you can call all the datasets used in the book.

Probability theory and statistics review

Basic probability theory We begin with some definitions. A **random process** is a process that can be repeated many times with different outcomes. The **sample space** is the set of all possible outcomes of a random process. We distinguish between **discrete** and **continuous** random process in the following table.

Description	Type	Potential outcomes
Coin	Discrete	Heads, Tails
6-sided die	Discrete	1, 2, 3, 4, 5, 6
Deck of cards	Discrete	2 ♦, 3 ♦, ... King ♦, Ace ♦
Housing prices	Continuous	$P \geq 0$

Table 1: Examples of Discrete and Continuous Random Processes.

We define **independent events** two ways. The first definition refers to logical independence. For instance, two events occur but there is no reason to believe that two events affect one another. The logical fallacy is called *post hoc ergo propter hoc*. The second definition is statistical independence. We'll illustrate the latter with an example from the idea of sampling with and without replacement. Let's use a randomly shuffled deck of cards for example. For a deck of 52 cards, what is the probability that the first card will be an Ace?

$$Pr(Ace) = \frac{\text{Count Aces}}{\text{Sample Space}} = \frac{4}{52} = \frac{1}{13} = 0.077$$

There are 52 possible outcomes, which is the sample space – it is the set of all possible outcomes of the random process. Of those 52 possible outcomes, we are concerned with the frequency of an Ace occurring. There are four Aces in the deck, so $\frac{4}{52} = 0.077$.

Assume that the first card was an ace. Now we ask the question again. If we shuffle the deck, what is the probability the next card drawn is an Ace? It is no longer $\frac{1}{13}$ because we did not “sample with replacement”. We sampled *without* replacement. Thus the new probability is

$$Pr(Ace | Card 1 = Ace) = \frac{3}{51} = 0.059$$

Under sampling without replacement, the two events – Ace on Card 1 and an Ace on Card 2 if Card 1 was an Ace – aren't independent events. To make the two events independent, you would have to put the Ace back and shuffle the deck. So two events, A and B , are independent if and only if (iff):

$$Pr(A|B) = Pr(A)$$

An example of two independent events would be rolling a 5 with one die after having rolled a 3 with another die. The two events are independent, so the probability of rolling a 5 is always 0.17 regardless of what we rolled on the first die.¹⁰

But what if we are wanting to know the probability of some event occurring that requires multiple events, first, to occur? For instance, let's say we're talking about the Cleveland Cavaliers winning the NBA championship. In 2016, the Golden State Warriors were 3-0 in a best of seven playoff. What had to happen for the Warriors to lose the playoff? The Cavaliers had to win four in a row. In this instance, to find the probability, we have to take the product of all marginal probabilities, or $Pr(\cdot)^n$ where $Pr(\cdot)$ is the marginal probability of one event occurring, and n is the number of repetitions of that one event. If the probability of each win is 0.5, and each game is independent, then it is the *product* of each game's probability of winning:

$$\text{Win probability} = Pr(W, W, W, W) = (0.5)^4 = 0.0625$$

Another example may be helpful. Let's say a batter has a 0.3 probability of getting a hit. Then what is the probability of him getting two hits in a row? The two hit probability is $Pr(HH) = 0.3^2 = 0.09$ and the three hit probability is $Pr(HHH) = 0.3^3 = 0.027$. Or to keep with our poker example, what is the probability of being dealt pocket Aces? It's $\frac{4}{52} \times \frac{3}{51} = 0.0045$ or 0.45%.

Let's now formalize what we've been saying for a more generalized case. For independent events, calculating *joint probabilities* is to multiply the marginal probabilities:

$$Pr(A, B) = Pr(A)Pr(B)$$

where $Pr(A, B)$ is the joint probability of both A and B occurring, and $Pr(A)$ is the marginal probability of A event occurring.

Now for a slightly more difficult application. What is the probability of rolling a 7 using two dice, and is it the same as the probability of rolling a 3? To answer this, let's compare the two. We'll use a table to help explain the intuition. First, let's look at all the ways to get a 7 using two six-sided die. There are 36 total possible outcomes ($6^2 = 36$) when rolling two dice. In Table 2 we see that

¹⁰ The probability rolling a 5 using one die is $\frac{1}{6} = 0.167$.

there are six different ways to roll a 7 using only two dice. So the probability of rolling a 7 is $6/36 = 16.67\%$. Next, let's look at all the ways to get a 3 using two six-sided dice. Table 3 shows that there are only two ways to get a 3 rolling two six-sided dice. So the probability of rolling a 3 is $2/36 = 5.56\%$. So, no, the probabilities are different.

Die 1	Die 2	Outcome
1	6	7
2	5	7
3	4	7
4	3	7
5	2	7
6	1	7

Table 2: Total number of ways to get a 7 with two six-sided dice.

Die 1	Die 2	Outcome
1	2	3
2	1	3

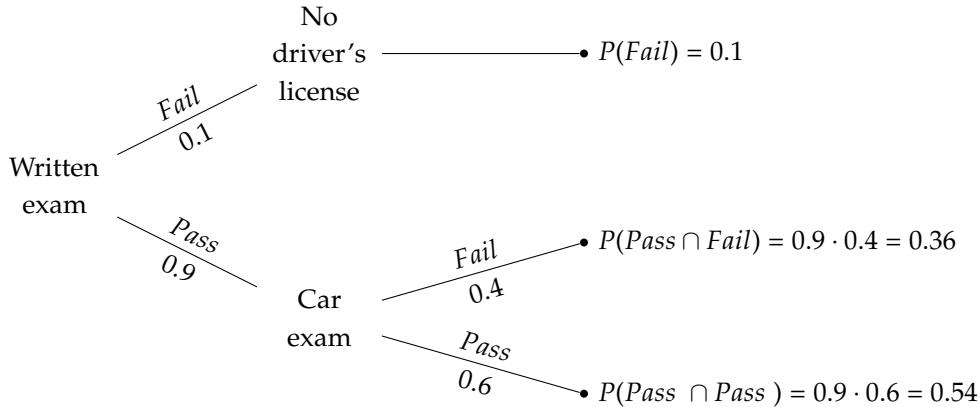
Table 3: Total number of ways to get a 3 using two six-sided dice.

Events and Conditional Probability First, before we talk about the three ways of representing a probability, I'd like to introduce some new terminology and concepts: *events* and *conditional probabilities*. Let A be some event. And let B be some other event. For two events, there are four possibilities.

1. A and B: Both A and B occur.
2. $\sim A$ and B: A does not occur, but B occurs.
3. A and $\sim B$: A occurs but B does not occur.
4. $\sim A$ and $\sim B$: Neither A nor B occurs.

I'll use a couple of different examples to illustrate how to represent a probability.

Probability tree Let's think about a situation where you are trying to get your driver's license. Suppose that in order to get a driver's license, you have to pass the written and the driving exam. However, if you fail the written exam, you're not allowed to take the driving exam. We can represent these two events in a probability tree.

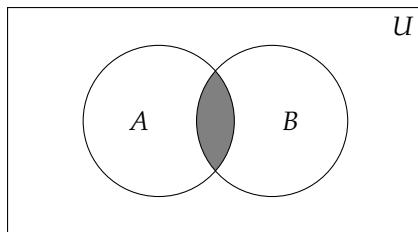


Probability trees are intuitive and easy to interpret. First, we see that the probability of passing the exam is 0.9 and the probability of failing the exam is 0.1. Second, at every branching off from a node, we can further see that the probabilities are summing to 1.0. For example, the probability of failing the written exam (0.1) plus the probability of passing it (0.9) equals 1.0. The probability of failing the car exam (0.4) plus the probability of passing the car exam (0.6) is 1.0. And finally, the joint probabilities are also all summing to 1.0. This is called the **law of total probability** is equal to the sum of all joint probability of A and B_n events occurring.

$$Pr(A) = \sum_n Pr(A \cap B_n)$$

We also see the concept of a conditional probability in this tree. For instance, the probability of failing the car exam conditional on passing the written exam is represented as $Pr(Fail|Pass) = 0.4$.

Venn Diagram A second way to represent multiple events occurring is with a Venn diagram. Venn diagrams were first conceived by John Venn in 1880 and are used to teach elementary set theory, as well as express set relationships in probability and statistics. This example will involve two sets, A and B .



Let's return to our earlier example of your team making the playoffs, which determines whether your coach is rehired. Here we remind ourselves of our terms. A and B are events, and U is the universal set of which A and B are subsets. Let A be the probability that

your team makes the playoffs and B is the probability your coach is rehired. Let $Pr(A) = 0.6$ and let $Pr(B) = 0.8$. Let the probability that both A and B occur be $Pr(A, B) = 0.5$.

Note, that $A + \sim A = U$, where $\sim A$ is the complement of A . The **complement** means that it is everything in the universal set that is not A . The same is said of B . The sum of B and $\sim B = U$. Therefore:

$$A + \sim A = B + \sim B$$

We can rewrite out the following definitions:

$$B = A + \sim A - \sim B$$

$$A = B + \sim B - \sim A$$

Additionally, whenever we want to describe a set of events in which either A or B could occur, it is: $A \cup B$.

So, here again we can also describe the shaded region as the union of the $\sim A \cup \sim B$ sets. Where $\sim A$ and $\sim B$ occurs is the outer area of the $A \cup B$. So again,

$$A \cap B + \sim A \cap \sim B = 1$$

Finally, the joint sets. These are those subsets wherein *both* A and B occur. Using the set notation, we can calculate the probabilities because for a Venn diagram with the overlapping circles (events), there are four possible outcomes expressed as joint sets.

Notice these relationships

$$A \cup B = A \cap \sim B + A \cap B + \sim A \cap B$$

$$A = A \cap \sim B + A \cap B$$

$$B = A \cap B + \sim A \cap B$$

Now it is just simple addition to find all missing values. Recall the A is your team making playoffs and $Pr(A) = 0.6$. And B is the probability the coach is rehired, $Pr(B) = 0.8$. Also, $Pr(A, B) = 0.5$ which is the probability of both A and B occurring. Then we have:

$$A = A \cap \sim B + A \cap B$$

$$A \cap \sim B = A - A \cap B$$

$$Pr(A, \sim B) = Pr(A) - Pr(A, B)$$

$$Pr(A, \sim B) = 0.6 - 0.5$$

$$Pr(A, \sim B) = 0.1$$

When working with set objects, it is important to understand that probabilities should be measured by considering the share of the larger subset, say A , that some subset takes, such as $A \cap B$. When

we write down that the probability of $A \cap B$ occurs at all, it is with regards to U . But what if we were to ask the question as “What share of A is due to $A \cap B$?” Notice, then, that we would need to do this:

$$? = A \cap B \div A$$

$$? = 0.5 \div 0.6$$

$$? = 0.83$$

I left this intentionally undefined on the left hand side so as to focus on the calculation itself. But now let’s define what we are wanting to calculate. “In a world where A has occurred, what is the probability B will also occur?” This is:

$$\begin{aligned} \text{Prob}(B \mid A) &= \frac{\text{Pr}(A, B)}{\text{Pr}(A)} = \frac{0.5}{0.6} = 0.83 \\ \text{Prob}(A \mid B) &= \frac{\text{Pr}(A, B)}{\text{Pr}(B)} = \frac{0.5}{0.8} = 0.63 \end{aligned}$$

Notice, these conditional probabilities are not as easily seen in the Venn diagram. We are essentially asking what percent of a subset – e.g., $\text{Pr}(A)$ – is due to the joint, for example $\text{Pr}(A, B)$. That is the notion of the conditional probability.

Contingency tables The last way that we can represent events is with a contingency table. Contingency tables are also sometimes called two way tables. Table 4 is an example of a contingency table. We continue our example about the coach.

Event labels	Coach is not rehired (B)	Coach is rehired ($\sim B$)	Total
(A) team playoffs	$\text{Pr}(A, \sim B)=0.1$	$\text{Pr}(A, B)=0.5$	$\text{Pr}(A)=0.6$
($\sim A$) no playoffs	$\text{Pr}(\sim A, \sim B)=0.1$	$\text{Pr}(\sim A, B)=0.3$	$\text{Pr}(B)=0.4$
Total	$\text{Pr}(\sim B)=0.2$	$\text{Pr}(B)=0.8$	1.0

Table 4: Two way contingency table.

Recall that $\text{Pr}(A)=0.6$, $\text{Pr}(B)=0.8$, and $\text{Pr}(A, B)=0.5$. All probabilities must sum correctly. Note that to calculate conditional probabilities, we must ask the frequency of the element in question (e.g., $\text{Pr}(A, B)$) relative to some other larger event (e.g., $\text{Pr}(A)$). So if we want to ask, “what is the conditional probability of B given A ?”, it’s:

$$\text{Pr}(B \mid A) = \frac{\text{Pr}(A, B)}{\text{Pr}(A)} = \frac{0.5}{0.6} = 0.83$$

but note to ask the frequency of $A \cup B$ in a world where B occurs is to ask the following:

$$\text{Pr}(A \mid B) = \frac{\text{Pr}(A, B)}{\text{Pr}(B)} = \frac{0.5}{0.8} = 0.63$$

So, we can use what we have done so far to write out a definition of joint probability. Let's start with a definition of conditional probability first. Given two events, A and B :

$$Pr(A|B) = \frac{Pr(A, B)}{Pr(B)} \quad (1)$$

$$Pr(B|A) = \frac{Pr(B, A)}{Pr(A)} \quad (2)$$

$$Pr(A, B) = Pr(B, A) \quad (3)$$

$$Pr(A) = Pr(A, \sim B) + Pr(\sim A, B) \quad (4)$$

$$Pr(B) = Pr(A, B) + Pr(\sim A, B) \quad (5)$$

Using equations 1 and 2, I can simply write down a definition of joint probabilities.

$$Pr(A, B) = Pr(A|B)Pr(B) \quad (6)$$

$$Pr(B, A) = Pr(B|A)Pr(A) \quad (7)$$

And this is simply the formula for joint probability. Given equation 3, and using the definitions of $Pr(A, B)$ and $Pr(B, A)$, I can also rearrange terms, make a substitution and rewrite it as:

$$\begin{aligned} Pr(A|B)Pr(B) &= Pr(B|A)Pr(A) \\ Pr(A|B) &= \frac{Pr(B|A)Pr(A)}{Pr(B)} \end{aligned} \quad (8)$$

Equation 8 is sometimes called the naive version of Bayes Rule. We will now decompose this more fully, though, by substituting equation 5 into equation 8.

$$Pr(A|B) = \frac{Pr(B|A)Pr(A)}{Pr(A, B) + Pr(\sim A, B)} \quad (9)$$

Substituting equation 6 into the denominator for equation 9 yields:

$$Pr(A|B) = \frac{Pr(B|A)Pr(A)}{Pr(B|A)Pr(A) + Pr(\sim A, B)} \quad (10)$$

Finally, we note that using the definition of joint probability, that $Pr(B, \sim A) = Pr(B|\sim A)Pr(\sim A)$, which we substitute into the denominator of equation 10 to get:

$$Pr(A|B) = \frac{Pr(B|A)Pr(A)}{Pr(B|A)Pr(A) + Pr(B|\sim A)Pr(\sim A)} \quad (11)$$

That's a mouthful of substitutions so what does equation 11 mean? This is the Bayesian decomposition version of Bayes rule. Let's use our example again of a team making the playoffs. A is your team makes the playoffs and B is your coach gets rehired. And $A \cap B$ is the

joint probability that both events occur. We can make each calculation using the contingency tables. The questions here is “if coach is rehired, what’s the probability that my team made the playoffs?” Or formally, $Pr(A|B)$. We can use the Bayesian decomposition to find what this equals.

$$\begin{aligned} Pr(A|B) &= \frac{Pr(B|A)Pr(A)}{Pr(B|A)Pr(A) + Pr(B|\sim A)Pr(\sim A)} \\ &= \frac{0.83 \cdot 0.6}{0.83 \cdot 0.6 + 0.75 \cdot 0.4} \\ &= \frac{0.498}{0.498 + 0.3} \\ &= \frac{0.498}{0.798} \\ Pr(A|B) &= 0.624 \end{aligned}$$

Check this against the contingency table using the definition of joint probability:

$$Pr(A|B) = \frac{Pr(A, B)}{Pr(B)} = \frac{0.5}{0.8} = 0.625$$

Why are they different? Because 0.83 is an approximation of $Pr(B|A)$ which was technically 0.833... trailing. So, if my coach is rehired, there is a 63 percent chance we will win.

Monty Hall example Let’s use a different example. This is a fun one, because most people find it counterintuitive. It even used to stump mathematicians and statisticians.¹¹ But Bayes rule makes the answer very clear.

Let’s assume three closed doors: Door 1 (D_1), Door 2 (D_2) and Door 3 (D_3). Behind one of the doors is a million dollars. Behind each of the other two doors is a goat. Monty Hall, the game show host in this example, asks the contestants to pick a door. After they had picked the door, but before he opens their door, he opens one of the other two doors to reveal a goat. He then ask the contestant, “would you like to switch doors?”

Many people answer say it makes no sense to change doors, because (they say) there’s an equal chance that the million dollars is behind either door. Therefore, why switch? There’s a 50/50 chance it’s behind the door picked and there’s a 50/50 chance it’s behind the remaining door, so it makes no rational sense to switch. Right? Yet, a little intuition should tell you that’s not the right answer, because it would seem that when Monty Hall opened that third door, he told us *something*. But what did he tell us exactly?

Let’s formalize the problem using our probability notation. Assume that you chose door 1, D_1 . What was the probability that

¹¹ There’s a fun story in which someone posed this question to the columnist, Marilyn vos Savant, and she got it right. People wrote in, calling her stupid, but it turns out she was right. You can read the story [here](#).

D_1 had a million dollars when you made that choice? $Pr(D_1 = 1 \text{ million}) = \frac{1}{3}$. We will call that event A_1 . And the probability that D_1 has a million dollars at the start of the game is $\frac{1}{3}$ because the sample space is 3 doors, of which one has a million dollars. Thus, $Pr(A_1) = \frac{1}{3}$. Also, by the law of total probability, $Pr(\sim A_1) = \frac{2}{3}$. Let's say that Monty Hall had opened door 2, D_2 , to reveal a goat. Then he then asked you "would you like to change to door number 3?"

We need to know the probability that door 3 has the million dollars and compare that to Door 1's probability. We will call the opening of door 2 event B . We will call the probability that the million dollars is behind door i , A_i . We now write out the question just asked formally and decompose it using the Bayesian decomposition. We are ultimately interested in knowing, "what is the probability that Door 1 has a million dollars (event A_1) given Monty Hall opened Door 2 (event B)", which is a conditional probability question. Let's write out that conditional probability now using the Bayesian decomposition from equation 11.

$$Pr(A_1|B) = \frac{Pr(B|A_1)Pr(A_1)}{Pr(B|A_1)Pr(A_1) + Pr(B|A_2)Pr(A_2) + Pr(B|A_3)Pr(A_3)} \quad (12)$$

There's basically two kinds of probabilities on the right-hand-side. There's the marginal probability that the million dollars is behind a given door $Pr(A_i)$. And there's the conditional probability that Monty Hall would open Door 2 given the million dollars is behind Door A_i , $Pr(B|A_i)$.

The marginal probability that Door i has the million dollars without any additional information is $\frac{1}{3}$. We call this the *prior* probability, or *prior belief*. It may also be called the *unconditional* probability.

The conditional probability, $Pr(B|A_i)$, require a little more careful thinking. Take the first conditional probability, $Pr(B|A_1)$. In a world where Door 1 has the million dollars, what's the probability Monty Hall would open door number 2? Think about it for a second.

Let's think about the second conditional probability: $Pr(B|A_2)$. In a world where the money is behind Door 2, what's the probability that Monty Hall would open Door 2? Think about it, too, for a second.

And then the last conditional probability, $Pr(B|A_3)$. In a world where the money is behind Door 3, what's the probability Monty Hall will open Door 2?

Each of these conditional probabilities require thinking carefully about the feasibility of the events in question. Let's examine the easiest question: $Pr(B|A_2)$. In a world where the money is behind Door 2, how likely is it for Monty Hall to open that same door, Door 2? Keep in mind: this is a game show. So that gives you some idea about how the game show host will behave. Do you think Monty

Hall would open a door that had the million dollars behind it? After all, isn't he opening doors that don't have it to ask you whether you should switch? It makes no sense to think he'd ever open a door that actually had the money behind it. Historically, even, he always opens a door with a goat. So don't you think he's only opening doors with goats? Let's see what happens if we take that intuition to its logical extreme and conclude that Monty Hall *never* opens a door if it has a million dollars. He *only* opens doors if those doors have a goat. Under that assumption, we can proceed to estimate $Pr(A_1|B)$ by substituting values for $Pr(B|A_i)$ and $Pr(A_i)$ into the right-hand-side of equation 12.

What then is $Pr(B|A_1)$? That is, in a world where *you* have chosen Door 1, and the money is behind Door 1, what is the probability that he would open Door 2? There are two doors he could open if the money is behind Door 1 – he could open either Door 2 or Door 3, as both have a goat. So $Pr(B|A_1) = 0.5$.

What about the second conditional probability, $Pr(B|A_2)$? In a world where the money is behind Door 2, what's the probability he will open it? Under our assumption that he never opens the door if it has a million dollars, we know this probability is 0.0. And finally, what about the third probability, $Pr(B|A_3)$? What is the probability he opens Door 2 given the money is behind Door 3? Now consider this one carefully - you have already chosen Door 1, so he can't open that one. And he can't open Door 3, because that has the money. The only door, therefore, he could open is Door 2. Thus, this probability is 1.0. And all the $Pr(A_i) = \frac{1}{3}$, allowing us to solve for the conditional probability on the left hand side through substitution, multiplication and division.

$$\begin{aligned} Pr(A_1|B) &= \frac{\frac{1}{2} \cdot \frac{1}{3}}{\frac{1}{2} \cdot \frac{1}{3} + 0 \cdot \frac{1}{3} + 1.0 \cdot \frac{1}{3}} \\ &= \frac{\frac{1}{6}}{\frac{1}{6} + \frac{2}{6}} \\ &= \frac{1}{3} \end{aligned}$$

Now, let's consider the probability that Door 3 has the million dollars, as that's the only door we can switch to since door two has been opened for us already.

$$\begin{aligned} Pr(A_3|B) &= \frac{Pr(B|A_3)Pr(A_3)}{Pr(B|A_3)Pr(A_3) + Pr(B|A_2)Pr(A_2) + Pr(B|A_1)Pr(A_1)} \\ &= \frac{1.0 \cdot \frac{1}{3}}{1.0 \cdot \frac{1}{3} + 0 \cdot \frac{1}{3} + \frac{1}{2} \cdot \frac{1}{3}} \\ &= \frac{2}{3} \end{aligned}$$

Ah hah. Now isn't that just a little bit surprising? The probability you are holding the correct door is $\frac{1}{3}$ just as it was before Monty Hall opened Door 2. But now the probability that it is in Door 2 has increased from its prior probability. The prior probability, $Pr(A_3) = \frac{1}{3}$, changed to a new conditional probability value, $Pr(A_3|B) = \frac{2}{3}$. This new conditional probability is called the *posterior* probability, or *posterior* belief.¹² Given the information you learned from witnessing B , we correctly updated our beliefs about the likelihood that Door 3 had the million dollars.

Exercises Driving while intoxicated is defined as operating a motor vehicle with a blood alcohol content (BAC) at or above 0.08%. Standardized field sobriety tests (SFSTs) are often used as tools by officers in the field to determine if an arrest followed by a breath test is justified. However, breath tests are often not available in court for a variety of reasons, and under those circumstances, the SFSTs are frequently used as an indication of impairment and sometimes as an indicator that the subject has a $BAC \geq 0.08\%$.

Stuster and Burns [1998] conducted an experiment to estimate the accuracy of SFSTs. Seven San Diego Police Department officers administered STSTs on those stopped for suspicion of driving under the influence of alcohol. The officers were then instructed to carry out the SFSTs on the subjects, and then to note an estimated BAC based *only* on the SFST results.¹³ Subjects driving appropriately were not stopped or tested. However, "poor drivers" were included because they attracted the attention of the officers.¹⁴ The officers were asked to estimate the BAC values using SFSTs only. Some of the subjects were arrested and given a breath test. The criteria used by the officers for estimation of BAC was not described in the Stuster and Burns [1998] study, and several studies have concluded that officers were using the SFSTs to then subjectively guess at the subjects' BAC. There were 297 subjects in the original data. The raw data is reproduced below.

	MBAC < 0.08%	MBAC $\geq 0.08\%$
EBAC $\geq 0.08\%$	$n = 24$	$n = 210$
EBAC < 0.08%	$n = 59$	$n = 4$

¹² It's okay to giggle every time you say "posterior". I do!

¹³ In case you're interested, the SFST consists of three tests: the walk and turn test, the one leg stand test, and the horizontal gaze nystagmus test.

¹⁴ The data collected included gender and age, but not race, body weight, presence of prior injuries and other factors that might influence SFSTs of the measured BAC.

1. Represent the events above using a probability tree, two way tables, and a Venn Diagram. Calculate the marginal probability of each event, the joint probability, and the conditional probability.
2. Let F be the event where a driver fails the SFST with an estimated BAC (EBAC) at or above 0.08, and $\sim F$ be an event where the

driver passes ($EBAC < 0.08$). Let I be the event wherein a driver is impaired by alcohol with an *actual* or measured BAC (MBAC) is at or above 0.08%, and $\sim I$ be an event where $MBAC < 0.08$. Use Bayes Rule to decompose the conditional probability, $Pr(I|F)$, into it the correct expression. Label the prior beliefs, posterior beliefs, false and true positive, false and true negative if and where they apply. Show that the posterior belief calculated using Bayes Rule is equal to the value that could be directly calculated using the sample information above. Interpret the posterior belief in plain language. What does it mean?

3. Assume that because of concerns about profiling, a new policy is enacted. Police must randomly pull over all automobiles for suspected driving while intoxicated and apply the SFST to all drivers. Using survey information, such as from Gallup or some other reputable survey, what percent of the US population drinks and drives? If the test statistics from the sample are correct, then how likely is it that someone who fails the SFST is impaired under this new policy?

Properties of Regression

"I like cool people, but why should I care?
Because I'm busy tryna fit a line with the least squares"
– J-Wong

Summation operator Now we move on to a review of the least squares estimator.¹⁵ Before we begin, let's introduce some new notation starting with the **summation operator**. The Greek letter Σ (the capital Sigma) denotes the summation operator. Let x_1, x_2, \dots, x_n be a sequence of numbers. We can compactly write a sum of numbers using the summation operator as:

$$\sum_{i=1}^n x_i \equiv x_1 + x_2 + \dots + x_n$$

The letter i is called the index of summation. Other letters are sometimes used, such as j or k , as indices of summation. The subscripted variable simply represents a specific value of a random variable, x . The numbers 1 and n are the lower limit and upper limit of the summation. The expression $\sum_{i=1}^n$ can be stated in words as "sum the numbers x_i for all values of i from 1 to n ". An example can help clarify:

$$\sum_{i=6}^9 x_i = x_6 + x_7 + x_8 + x_9$$

The summation operator has three properties. The first property is called the constant rule. Formally, it is:

$$\text{For any constant } c : \sum_{i=1}^n c = nc \quad (13)$$

Let's consider an example. Say that we are given:

$$\sum_{i=1}^3 5 = (5 + 5 + 5) = 3 \cdot 5 = 15$$

A second property of the summation operator is:

$$\sum_{i=1}^n cx_i = c \sum_{i=1}^n x_i \quad (14)$$

¹⁵ This chapter is heavily drawn from Wooldridge [2010] and Wooldridge [2015]. All errors are my own.

Again let's use an example. Say we are given:

$$\begin{aligned}\sum_{i=1}^3 5x_i &= 5x_1 + 5x_2 + 5x_3 \\ &= 5(x_1 + x_2 + x_3) \\ &= 5 \sum_{i=1}^3 x_i\end{aligned}$$

We can apply both of these properties to get the following third property:

$$\text{For any constant } a \text{ and } b : \sum_{i=1}^n (ax_i + by_i) = a \sum_{i=1}^n x_i + b \sum_{j=1}^n y_j$$

Before leaving the summation operator, it is useful to also note things which are **not** properties of this operator. Two things which summation operators **cannot** do:

$$\begin{aligned}\sum_i \frac{x_i}{y_i} &\neq \frac{\sum_{i=1}^n x_i}{\sum_{i=1}^n y_i} \\ \sum_{i=1}^n x_i^2 &\neq \left(\sum_{i=1}^n x_i \right)^2\end{aligned}$$

We can use the summation indicator to make a number of calculations, some of which we will do repeatedly over the course of this book. For instance, we can use the summation operator to calculate the **average**:

$$\begin{aligned}\bar{x} &= \frac{1}{n} \sum_{i=1}^n x_i \\ &= \frac{x_1 + x_2 + \dots + x_n}{n}\end{aligned}\tag{15}$$

where \bar{x} is the average (mean) of the random variable x_i . Another calculation we can make is a random variable's deviations from its own mean. The sum of the deviations from the mean is always equal to zero:

$$\sum_{i=1}^n (x_i - \bar{x}) = 0\tag{16}$$

Let's illustrate this with an example in Table 5:

Consider a sequence of two numbers $\{y_1, y_2, \dots, y_n\}$ and $\{x_1, x_2, \dots, x_n\}$. Then we may consider double summations over possible values of x 's and y 's. For example, consider the case where $n = m = 2$. Then,

x	$x - \bar{x}$
10	2
4	-4
13	5
5	-3
Mean=8	Sum=0

Table 5: Sum of deviations equalling zero

$\sum_{i=1}^2 \sum_{j=1}^2 x_i y_j$ is equal to $x_1 y_1 + x_1 y_2 + x_2 y_1 + x_2 y_2$. This is because:

$$\begin{aligned}
x_1 y_1 + x_1 y_2 + x_2 y_1 + x_2 y_2 &= x_1(y_1 + y_2) + x_2(y_1 + y_2) \\
&= \sum_{i=1}^2 x_i(y_1 + y_2) \\
&= \sum_{i=1}^2 x_i \left(\sum_{j=1}^2 y_j \right) \\
&= \sum_{i=1}^2 \left(\sum_{j=1}^2 x_i y_j \right) \\
&= \sum_{i=1}^2 \sum_{j=1}^2 x_i y_j
\end{aligned}$$

One result that will be very useful throughout the semester is:

$$\sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n(\bar{x})^2 \quad (17)$$

An overly long, step-by-step, proof is below. Note that the summation index is suppressed after the first line for brevity sake.

$$\begin{aligned}
\sum_{i=1}^n (x_i - \bar{x})^2 &= \sum_{i=1}^n (x_i^2 - 2x_i \bar{x} + \bar{x}^2) \\
&= \sum x_i^2 - 2\bar{x} \sum x_i + n\bar{x}^2 \\
&= \sum x_i^2 - 2\frac{1}{n} \sum x_i \sum x_i + n\bar{x}^2 \\
&= \sum x_i^2 + n\bar{x}^2 - \frac{2}{n} \left(\sum x_i \right)^2 \\
&= \sum x_i^2 + n \left(\frac{1}{n} \sum x_i \right)^2 - 2n \left(\frac{1}{n} \sum x_i \right)^2 \\
&= \sum x_i^2 - n \left(\frac{1}{n} \sum x_i \right)^2 \\
&= \sum x_i^2 - n\bar{x}^2
\end{aligned}$$

A more general version of this result is:

$$\begin{aligned}
 \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) &= \sum_{i=1}^n x_i(y_i - \bar{y}) \\
 &= \sum_{i=1}^n (x_i - \bar{x})y_i \\
 &= \sum_{i=1}^n x_iy_i - n(\bar{x}\bar{y})
 \end{aligned} \tag{18}$$

Or:

$$\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i(y_i - \bar{y}) = \sum_{i=1}^n (x_i - \bar{x})y_i = \sum_{i=1}^n x_iy_i - n(\bar{x}\bar{y}) \tag{19}$$

Expected value The **expected value** of a random variable, also called the expectation and sometimes the population mean, is simply the weighted average of the possible values that the variable can take, with the weights being given by the probability of each value occurring in the population. Suppose that the variable X can take on values x_1, x_2, \dots, x_k each with probability $f(x_1), f(x_2), \dots, f(x_k)$, respectively. Then we define the expected value of X as:

$$\begin{aligned}
 E(X) &= x_1f(x_1) + x_2f(x_2) + \dots + x_kf(x_k) \\
 &= \sum_{j=1}^k x_jf(x_j)
 \end{aligned} \tag{20}$$

Let's look at a numerical example. If X takes on values of $-1, 0$ and 2 with probabilities $0.3, 0.3$ and 0.4 ,¹⁶ respectively. Then the expected value of X equals:

$$\begin{aligned}
 E(X) &= (-1)(0.3) + (0)(0.3) + (2)(0.4) \\
 &= 0.5
 \end{aligned}$$

In fact you could take the expectation of a function of that variable, too, such as X^2 . Note that X^2 takes only the values $1, 0$ and 4 with probabilities $0.3, 0.3$ and 0.4 . Calculating the expected value of X^2 therefore is:

$$\begin{aligned}
 E(X^2) &= (-1)^2(0.3) + (0)^2(0.3) + (2)^2(0.4) \\
 &= 1.9
 \end{aligned}$$

The first property of expected value is that for any constant, c , $E(c) = c$. The second property is that for any two constants, a and b , then $E(aX + b) = E(aX) + E(b) = aE(X) + b$. And the third property is that if we have numerous constants, a_1, \dots, a_n and many random variables, X_1, \dots, X_n , then the following is true:

$$E(a_1X_1 + \dots + a_nX_n) = a_1E(X_1) + \dots + a_nE(X_n)$$

¹⁶ Recall the law of total probability requires that all marginal probabilities sum to unity.

We can also express this using the expectation operator:

$$E\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i E(X_i)$$

And in the special case where $a_i = 1$, then

$$E\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n E(X_i)$$

Variance The expectation operator, $E(\cdot)$, is a **population** concept. It refers to the whole group of interest, not just the sample we have available to us. Its intuition is loosely similar to the average of a random variable in the population. Some additional properties for the expectation operator can be explained assuming two random variables, W and H .

$$\begin{aligned} E(aW + b) &= aE(W) + b \text{ for any constants } a, b \\ E(W + H) &= E(W) + E(H) \\ E(W - E(W)) &= 0 \end{aligned}$$

Consider the variance of a random variable, W :

$$V(W) = \sigma^2 = E[(W - E(W))^2] \text{ in the population}$$

We can show that:

$$V(W) = E(W^2) - E(W)^2 \quad (21)$$

In a given sample of data, we can estimate the variance by the following calculation:

$$\hat{S}^2 = (n - 1)^{-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

where we divide by $n - 1$ because we are making a degrees of freedom adjustment from estimating the mean. But in large samples, this degree of freedom adjustment has no practical effect on the value of \hat{S}^2 .¹⁷

A few more properties of variance. First, the variance of a line is:

$$V(aX + b) = a^2 V(X)$$

And the variance of a constant is zero (i.e., $V(c) = 0$ for any constant, c). The variance of the sum of two random variables is equal to:

$$V(X + Y) = V(X) + V(Y) + 2(E(XY) - E(X)E(Y)) \quad (22)$$

If the two variables are independent, then $E(XY) = E(X)E(Y)$ and $V(X + Y)$ is just equal to the sum of $V(X) + V(Y)$.

¹⁷ Whenever possible, I try to use the "hat" to represent an estimated statistic. Hence \hat{S}^2 instead of just S^2 . But it is probably more common to see the sample variance represented as S^2 .

Covariance The last part of equation 22 is called the covariance. The **covariance** measures the amount of linear dependence between two random variables. We represent it with the $C(X, Y)$ operator. $C(X, Y) > 0$ indicates that two variables move in the same direction, whereas $C(X, Y) < 0$ indicates they move in opposite directions. Thus we can rewrite Equation 22 as:

$$V(X + Y) = V(X) + V(Y) + 2C(X, Y)$$

While it's tempting to say that a zero covariance means two random variables are unrelated, that is incorrect. They could have a nonlinear relationship. The definition of covariance is

$$C(X, Y) = E(XY) - E(X)E(Y) \quad (23)$$

As we said, if X and Y are independent, then $C(X, Y) = 0$ in the population.¹⁸ The covariance between two linear functions is:

$$C(a_1 + b_1 X, a_2 + b_2 Y) = b_1 b_2 C(X, Y)$$

The two constants, a_1 and a_2 , zero out because their mean is themselves and so the difference equals zero.

Interpreting the magnitude of the covariance can be tricky. For that, we are better served looking at **correlation**. We define correlation as follows. Let $W = \frac{X - E(X)}{\sqrt{V(X)}}$ and $Z = \frac{Y - E(Y)}{\sqrt{V(Y)}}$. Then:

$$\text{Corr}(W, Z) = \frac{C(X, Y)}{\sqrt{V(X)V(Y)}} \quad (24)$$

The correlation coefficient is bounded between -1 and 1 . A positive (negative) correlation indicates that the variables move in the same (opposite) ways. The closer to 1 or -1 the stronger the linear relationship is.

Population model We begin with cross-sectional analysis. We will also assume that we can collect a random sample from the population of interest. Assume there are two variables, x and y , and we would like to see how y varies with changes in x .¹⁹

There are three questions that immediately come up. One, what if y is affected by factors other than x ? How will we handle that? Two, what is the functional form connecting these two variables? Three, if we are interested in the causal effect of x on y , then how can we distinguish that from mere correlation? Let's start with a specific model.

$$y = \beta_0 + \beta_1 x + u \quad (25)$$

This model is assumed to hold in the *population*. Equation 25 defines a **linear bivariate regression model**. For causal inference, the terms

¹⁸ It may be redundant to keep saying this, but since we've been talking about only the population this whole time, I wanted to stress it again for the reader.

¹⁹ Notice – this is not necessarily causal language. We are speaking first and generally just in terms of two random variables systematically moving together in some measurable way.

on the left-hand-side are usually thought of as the effect, and the terms on the right-hand-side are thought of as the causes.

Equation 25 explicitly allows for other factors to affect y by including a random variable called the **error term**, u . This equation also explicitly models the functional form by assuming that y is linearly dependent on x . We call the β_0 coefficient the **intercept parameter**, and we call the β_1 coefficient the **slope parameter**. These, note, describe a population, and our goal in empirical work is estimate their values. I will emphasize this several times throughout this book: we never directly observe these parameters, because they are not data. What we can do, though, is estimate these parameters using *data* and *assumptions*. We just have to have credible assumptions to *accurately* estimate these parameters with data. We will return to this point later. In this simple regression framework, all unobserved variables are subsumed by the error term.

First, we make a simplifying assumption without loss of generality. Let the expected value of u be zero in the population. Formally:

$$E(u) = 0 \quad (26)$$

where $E(\cdot)$ is the expected value operator discussed earlier. Normalizing ability to be zero in the population is harmless. Why? Because the presence of β_0 (the intercept term) always allows us this flexibility. If the average of u is different from zero – for instance, say that it's α_0 – then we just adjust the intercept. Adjusting the intercept has no effect on the β_1 slope parameter, though.

$$y = (\beta_0 + \alpha_0) + \beta_1 x + (u - \alpha_0)$$

where $\alpha_0 = E(u)$. The new error term is $u - \alpha_0$ and the new intercept term is $\beta_0 + \alpha_0$. But while those two terms changed, notice what did *not* change: the slope, β_1 , has not changed.

Mean independence An assumption that meshes well with our elementary treatment of statistics involves the mean of the error term for each “slice” of the population determined by values of x :

$$E(u|x) = E(u) \text{ for all values } x \quad (27)$$

where $E(u|x)$ means the “expected value of u given x ”. If equation 27 holds, then we say that u is **mean independent** of x . An example might help here. Let's say we are estimating the effect of schooling on wages, and u is unobserved ability. Mean independence requires that $E[ability|x = 8] = E[ability|x = 12] = E[ability|x = 16]$ so that the average ability is the same in the different portions of the population with an 8th grade education, a 12th grade education and a college

education. Because people choose education, though, based partly on that unobserved ability, equation 27 is almost certainly violated in this actual example.

Combining this new assumption, $E[u|x] = E[u]$ (a non-trivial assumption to make), with $E[u] = 0$ (a normalization and trivial assumption), and you get the following new assumption:

$$E(u|x) = 0, \text{ for all values } x \quad (28)$$

Equation 28 is called the **zero conditional mean assumption** and is a key identifying assumption in regression models. Because the conditional expected value is a linear operator, $E(u|x) = 0$ implies

$$E(y|x) = \beta_0 + \beta_1 x$$

which shows the **population regression function** is a linear function of x , or what Angrist and Pischke [2009] call the conditional expectation function.²⁰ This relationship is crucial for the intuition of the parameter, β_1 , as a *causal parameter*.

Least Squares Given data on x and y , how can we estimate the population parameters, β_0 and β_1 ? Let $\{(x_i, y_i) : i = 1, 2, \dots, n\}$ be a **random sample** of size n (the number of observations) from the population. Plug any observation into the population equation:

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

where i indicates a particular observation. We observe y_i and x_i but not u_i . We just know that u_i is there. We then use the two population restrictions that we discussed earlier:

$$\begin{aligned} E(u) &= 0 \\ C(x, u) &= 0 \end{aligned}$$

to obtain estimating equations for β_0 and β_1 . We talked about the first condition already. The second one, though, means that x and u are *uncorrelated* because recall covariance is the numerator of correlation equation (equation 24). Both of these conditions imply equation 28:

$$E(u|x) = 0$$

With $E(xu) = 0$, we get $E(u) = 0$, $C(x, u) = 0$. Notice that if $C(x, u) = 0$, it implies x and u are independent.²¹ Next we plug in for u , which is equal to $y - \beta_0 - \beta_1 x$:

$$\begin{aligned} E(y - \beta_0 - \beta_1 x) &= 0 \\ E(x[y - \beta_0 - \beta_1 x]) &= 0 \end{aligned}$$

²⁰ Notice that the conditional expectation passed through the linear function leaving a constant, because of the first property of the expectation operator, and a constant times x . This is because the conditional expectation of $E[X|X] = X$. This leaves us with $E[u|X]$ which under zero conditional mean is equal to zero.

²¹ See equation 23.

These are the two conditions in the **population** that effectively determine β_0 and β_1 . And again, note that the notation here is population concepts. We don't have access to populations, though we do have their sample counterparts:

$$\frac{1}{n} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \quad (29)$$

$$\frac{1}{n} \sum_{i=1}^n x_i [y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i] = 0 \quad (30)$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimates from the data.²² These are two linear equations in the two unknowns $\hat{\beta}_0$ and $\hat{\beta}_1$. Recall the properties of the summation operator as we work through the following sample properties of these two equations. We begin with equation 29 and pass the summation operator through.

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) &= \frac{1}{n} \sum_{i=1}^n (y_i) - \frac{1}{n} \sum_{i=1}^n \hat{\beta}_0 - \frac{1}{n} \sum_{i=1}^n \hat{\beta}_1 x_i \\ &= \frac{1}{n} \sum_{i=1}^n y_i - \hat{\beta}_0 - \hat{\beta}_1 \left(\frac{1}{n} \sum_{i=1}^n x_i \right) \\ &= \bar{y} - \hat{\beta}_0 - \hat{\beta}_1 \bar{x} \end{aligned}$$

where $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ which is the average of the n numbers $\{y_i : 1, \dots, n\}$. For emphasis we will call \bar{y} the **sample average**. We have already shown that the first equation equals zero (Equation 29), so this implies $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$. So we now use this equation to write the intercept in terms of the slope:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

We now plug $\hat{\beta}_0$ into the second equation, $\sum_{i=1}^n x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$. This gives us the following (with some simple algebraic manipulation):

$$\begin{aligned} \sum_{i=1}^n x_i [y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) - \hat{\beta}_1 x_i] &= 0 \\ \sum_{i=1}^n x_i (y_i - \bar{y}) &= \hat{\beta}_1 \left[\sum_{i=1}^n x_i (x_i - \bar{x}) \right] \end{aligned}$$

So the equation to solve is²³

$$\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \hat{\beta}_1 \left[\sum_{i=1}^n (x_i - \bar{x})^2 \right]$$

If $\sum_{i=1}^n (x_i - \bar{x})^2 \neq 0$, we can write:

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \frac{\text{Sample covariance}(x_i, y_i)}{\text{Sample variance}(x_i)} \end{aligned} \quad (31)$$

²² Notice that we are dividing by n , not $n - 1$. There is no degrees of freedom correction, in other words, when using samples to calculate means. There is a degrees of freedom correction when we start calculating higher moments.

²³ Recall from much earlier that:

$$\begin{aligned} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) &= \sum_{i=1}^n x_i (y_i - \bar{y}) \\ &= \sum_{i=1}^n (x_i - \bar{x}) y_i \\ &= \sum_{i=1}^n x_i y_i - n(\bar{x}\bar{y}) \end{aligned}$$

The previous formula for $\hat{\beta}_1$ is important because it shows us how to take data that we have and compute the slope estimate. The estimate, $\hat{\beta}_1$, is commonly referred to as the **ordinary least squares (OLS)** slope estimate. It can be computed whenever the sample variance of x_i isn't zero. In other words, if x_i is not constant across all values of i . The intuition is that the variation in x is what permits us to identify its impact in y . This also means, though, that we cannot determine the slope in a relationship if we observe a sample where everyone has the same years of schooling, or whatever causal variable we are interested in.

Once we have calculated $\hat{\beta}_1$, we can compute the intercept value, $\hat{\beta}_0$ as $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$. This is the OLS intercept *estimate* because it is calculated using sample averages. Notice that it is straightforward because $\hat{\beta}_0$ is linear in $\hat{\beta}_1$. With computers and statistical programming languages and software, we let our computers do these calculations because even when n is small, these calculations are quite tedious.²⁴

For any candidate estimates, $\hat{\beta}_0, \hat{\beta}_1$, we define a **fitted value** for each i as:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Recall that $i = \{1, \dots, n\}$ so we have n of these equations. This is the value we predict for y_i given that $x = x_i$. But there is prediction error because $y \neq \hat{y}_i$. We call that mistake the **residual**, and here use the \hat{u}_i notation for it. So the residual equals:

$$\begin{aligned}\hat{u}_i &= y_i - \hat{y}_i \\ \hat{u}_i &= y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\end{aligned}$$

Suppose we measure the size of the mistake, for each i , by squaring it. Squaring it will, after all, eliminate all negative values of the mistake so that everything is a positive value. This becomes useful when summing the mistakes if we aren't wanting positive and negative values to cancel one another out. So let's do that: square the mistake and add them all up to get, $\sum_{i=1}^n \hat{u}_i^2$:

$$\begin{aligned}\sum_{i=1}^n \hat{u}_i^2 &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2\end{aligned}$$

This equation is called the **sum of squared residuals** because the residual is $\hat{u}_i = y_i - \hat{y}_i$. But, the residual is based on estimates of the slope and the intercept. We can imagine any number of estimates of those values. But what if our goal is to *minimize* the sum of squared residuals by choosing $\hat{\beta}_0$ and $\hat{\beta}_1$? Using calculus, it can be shown that

²⁴ Back in the old days, though? Let's be glad that the old days of calculating OLS estimates by hand is long gone.

the solutions to that problem yields parameter estimates that are the same as what we obtained before.

Once we have the numbers $\hat{\beta}_0$ and $\hat{\beta}_1$ for a given dataset, we write the **OLS Regression line**:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \quad (32)$$

Let's consider an example in Stata.

```
set seed 1
clear
set obs 10000
gen x = rnormal()
gen u = rnormal()
gen y = 5.5*x + 12*u
reg y x
predict yhat1
gen yhat2 = 0.0732608 + 5.685033*x
sum yhat*
predict uhat1, residual
gen uhat2=y-yhat2
sum uhat*
twoway (lfit y x, lcolor(black) lwidth(medium)) (scatter
y x, mcolor(black) msymbol(point)), title(OLS
Regression Line)
rvfplot, yline(0)
```

Run the previous lines verbatim into Stata. Notice that the estimated coefficients – y-intercept and slope parameter – are represented in blue and red below in Figure 3.

Recall that we defined the fitted value as \hat{y}_i and we defined the residual, \hat{u}_i , as $y_i - \hat{y}_i$. Notice that the scatter plot relationship between the residuals and the fitted values created a spherical pattern suggesting that they are uncorrelated (Figure 4).

Once we have the estimated coefficients, and we have the OLS regression line, we can predict y (outcome) for any (sensible) value of x . So plug in certain values of x , we can immediately calculate what y will probably be with some error. The value of OLS here lies in how large that error is: OLS minimizes the error for a linear function. In fact, it is the best such guess at y for all linear estimators because it minimizes the prediction error. There's always prediction error, in other words, with any estimator, but OLS is the least worst.

Notice that the intercept is the predicted value of y if and when $x = 0$. Since here that value is 0.0732608, it's a little hard to read, but that's because x and u were random draws and so there's a value of

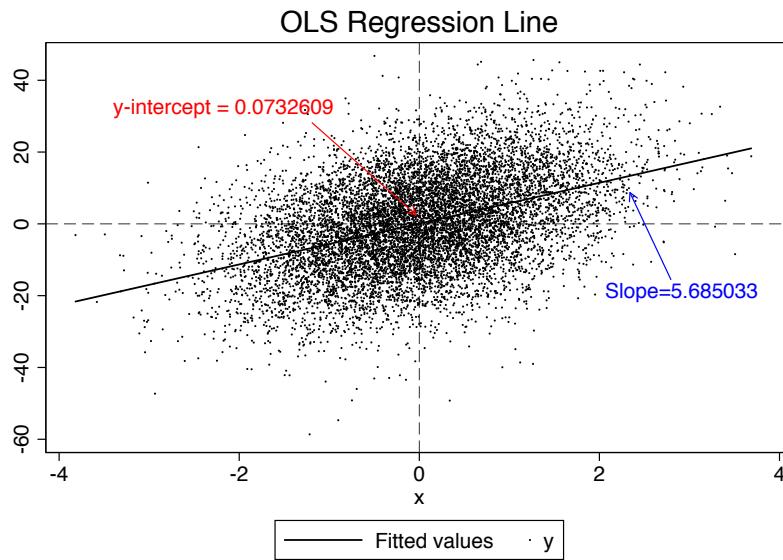


Figure 3: Graphical representation of bivariate regression from y on x

zero for y on average when $x = 0$.²⁵ The slope allows us to predict changes in y for any reasonable change in x according to:

$$\Delta \hat{y} = \hat{\beta}_1 \Delta x$$

And if $\Delta x = 1$, then x increases by one unit, and so $\Delta \hat{y} = 5.685033$ in our numerical example because $\hat{\beta}_1 = 5.685033$.

Now that we have calculated $\hat{\beta}_0$ and $\hat{\beta}_1$, we get the OLS fitted values by plugging the x_i into the following equation for $i = 1, \dots, n$:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

The OLS residuals are also calculated by:

$$\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

Most residuals will be different from zero (i.e., they do not lie on the regression line). You can see this in Figure 3. Most of the residuals are not on the regression line. Some are positive, and some are negative. A positive residual indicates that the regression line (and hence, the predicted values) underestimates the true value of y_i . And if the residual is negative, then it overestimated.

Algebraic Properties of OLS Remember how we obtained $\hat{\beta}_0$ and $\hat{\beta}_1$? When an intercept is included, we have:

$$\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

²⁵This is because on average u and x are independent, even if in the sample they aren't. Sample characteristics tend to be slightly different from population properties because of sampling error.

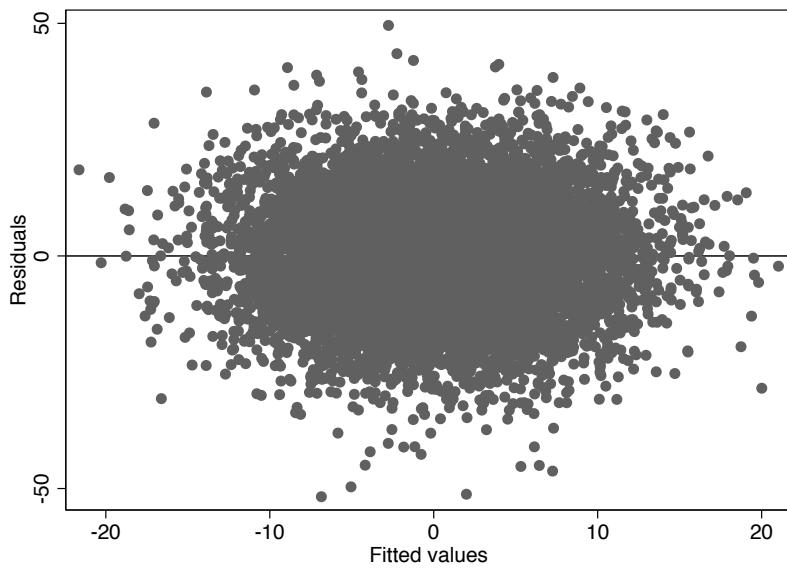


Figure 4: Distribution of residuals around regression line

The OLS residual *always* adds up to zero, by *construction*.

$$\sum_{i=1}^n \hat{u}_i = 0 \quad (33)$$

Sometimes seeing is believing, so let's look at this together. Type the following into Stata verbatim.

```
. clear
. set seed 1234
. set obs 10
. gen x = 9*rnorm()
. gen u = 36*rnorm()
. gen y = 3 + 2*x + u
. reg y x
. predict yhat
. predict residuals, residual
. su residuals
. list
. collapse (sum) x u y yhat residuals
. list
```

Output from this can be summarized in the following table (Table 6).

no.	x	u	y	\hat{y}	\hat{u}	$x\hat{u}$	$\hat{y}\hat{u}$
1.	-4.381653	-32.95803	-38.72134	-3.256034	-35.46531	155.3967	115.4762
2.	-13.28403	-8.028061	-31.59613	-26.30994	-5.28619	70.22192	139.0793
3.	-.0982034	17.80379	20.60738	7.836532	12.77085	-1.254141	100.0792
4.	-.1238423	-9.443188	-6.690872	7.770137	-14.46101	1.790884	-112.364
5.	4.640209	13.18046	25.40688	20.10728	5.353592	24.84179	107.6462
6.	-1.252096	-34.64874	-34.15294	4.848374	-39.00131	48.83337	-189.0929
7.	11.58586	9.118524	35.29023	38.09396	-2.80373	-32.48362	-106.8052
8.	-5.289957	82.23296	74.65305	-5.608207	80.26126	-424.5786	-450.1217
9.	-.2754041	11.60571	14.0549	7.377647	6.677258	-1.838944	49.26245
10.	-19.77159	-14.61257	-51.15575	-43.11034	-8.045414	159.0706	346.8405
Sum	-28.25072	34.25085	7.749418	7.749418	1.91e-06	-6.56e-06	.0000305

Notice the difference between the u , \hat{y} and \hat{u} columns. When we sum these ten lines, neither the error term nor the fitted values of y sum to zero. But the residuals *do sum to zero*. This is, as we said, one of the algebraic properties of OLS – coefficients were optimally chosen to ensure that the residuals sum to zero.

Because $y_i = \hat{y}_i + \hat{u}_i$ by definition (which we can also see in the above table), we can take the sample average of both sides

$$\frac{1}{n} \sum_{i=1}^n y_i = \frac{1}{n} \sum_{i=1}^n \hat{y}_i + \frac{1}{n} \sum_{i=1}^n \hat{u}_i$$

and so $\bar{y} = \bar{\hat{y}}$ because the residuals sum to zero. Similarly, the way that we obtained our estimates yields,

$$\sum_{i=1}^n x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

The sample covariance (and therefore the sample correlation) between the explanatory variables and the residuals is always zero (see Table 6):

$$\sum_{i=1}^n x_i \hat{u}_i = 0$$

Because the \hat{y}_i are linear functions of the x_i , the fitted values and residuals are uncorrelated too (See Table 6):

$$\sum_{i=1}^n \hat{y}_i \hat{u}_i = 0$$

Both properties hold by construction. In other words, $\hat{\beta}_0$ and $\hat{\beta}_1$ were selected to make them true.²⁶

A third property is that if we plug in the average for x , we predict the sample average for y . That is, the point, (\bar{x}, \bar{y}) is on the OLS regression line, or:

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$$

Table 6: Simulated data showing the sum of residuals equals zero

²⁶ Using the Stata code from Table 6, you can show all these algebraic properties yourself. I encourage you to do so by creating new variables equalling the product of these terms and collapsing as we did with the other variables. This will help you believe these algebraic properties hold.

Goodness of Fit For each observation, we write

$$y_i = \hat{y}_i + \hat{u}_i$$

Define the **total** (SST), **explained** (SSE) and **residual** (SSR) sum of squares as

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2 \quad (34)$$

$$SSE = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \quad (35)$$

$$SSR = \sum_{i=1}^n \hat{u}_i^2 \quad (36)$$

These are sample variances when divided by $n - 1$.²⁷ $\frac{SST}{n-1}$ is the sample variance of y_i , $\frac{SSE}{n-1}$ is the sample variance of \hat{y}_i , and $\frac{SSR}{n-1}$ is the sample variance of \hat{u}_i . With some simple manipulation rewrite equation 34:

$$\begin{aligned} SST &= \sum_{i=1}^n (y_i - \bar{y})^2 \\ &= \sum_{i=1}^n [(y_i - \hat{y}_i) - (\hat{y}_i - \bar{y})]^2 \\ &= \sum_{i=1}^n [\hat{u}_i - (\hat{y}_i - \bar{y})]^2 \end{aligned}$$

²⁷ Recall the earlier discussion about degrees of freedom correction.

And then using that the fitted values are uncorrelated with the residuals (equation 34), we can show that:

$$SST = SSE + SSR$$

Assuming $SST > 0$, we can define the fraction of the total variation in y_i that is explained by x_i (or the OLS regression line) as

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

which is called the **R-squared** of the regression. It can be shown to be equal to the *square* of the correlation between y_i and \hat{y}_i . Therefore $0 \leq R^2 \leq 1$. An R-squared of zero means no linear relationship between y_i and x_i and an R-squared of one means a perfect linear relationship (e.g., $y_i = x_i + 2$). As R^2 increases, the y_i are closer and closer to falling on the OLS regression line.

You don't want to fixate on R^2 in causal inference, though. It's a useful summary measure but it does not tell us about causality. Remember, we aren't trying to explain y ; we are trying to estimate causal effects. The R^2 tells us how much of the variation in y_i is explained by the explanatory variables. But if we are interested in the causal effect of a single variable, R^2 is irrelevant. For causal inference, we need equation 28.

Expected Value of OLS Up to now, we motivated simple regression using a population model. But our analysis has been purely algebraic based on a sample of data. So residuals always average to zero when we apply OLS to a sample, regardless of any underlying model. But now our job gets tougher. Now we have to study the statistical properties of the OLS estimator, referring to a population model and assuming random sampling.²⁸

Mathematical statistics is concerned with questions like “how do our estimators behave across different samples of data?” On average, for instance, will we get the right answer if we could repeatedly sample? We need to find the expected value of the OLS estimators – in effect the average outcome across all possible random samples – and determine if we are right on average. This leads naturally to a characteristic called **unbiasedness**, which is a desirable characteristic of all estimators.

$$E(\hat{\beta}) = \beta \quad (37)$$

Remember our objective is to estimate β_1 , which is the slope **population** parameter that describes the relationship between y and x . Our estimate, $\hat{\beta}_1$ is an **estimator** of that parameter obtained for a specific sample. Different samples will generate different estimates ($\hat{\beta}_1$) for the “true” (and unobserved) β_1 . Unbiasedness is the idea that if we could take as many random samples on Y as we want from the population and compute an estimate each time, the average of these estimates would be equal to β_1 .

There are several assumptions required for OLS to be unbiased. We will review those now. The first assumption is called “linear in the parameters”. Assume a population model of:

$$y = \beta_0 + \beta_1 x + u$$

where β_0 and β_1 are the unknown population parameters. We view x and u as outcomes of random variables generated by some data generating process. Thus, since y is a function of x and u , both of which are random, then y is also random. Stating this assumption formally shows our goal is to estimate β_0 and β_1 .

Our second assumption is “random sampling”. We have a random sample of size n , $\{(x_i, y_i) : i = 1, \dots, n\}$, following the population model. We know how to use this data to estimate β_0 and β_1 by OLS. Because each i is a draw from the population, we can write, for each i :

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

Notice that u_i here is the unobserved error for observation i . It is *not* the residual that we compute from the data.

²⁸ This section is a review of a traditional econometrics pedagogy. We cover it for the sake of completeness, as traditionally, econometricians motivated their discuss of causality through ideas like unbiasedness and consistency.

The third assumption is called the “sample variation in the explanatory variable”. That is, the sample outcomes on x_i are not all the same value. This is the same as saying the sample variance of x is not zero. In practice, this is no assumption at all. If the x_i are all the same value (i.e., constant), we cannot learn how x affects y in the population. Recall that OLS is the covariance of y and x divided by the variance in x and so if x is constant, then we are dividing by zero, and the OLS estimator is undefined.

The fourth assumption is where our assumptions start to have real teeth. It is called the “zero conditional mean” assumption and is probably the most critical assumption in causal inference. In the population, the error term has zero mean given any value of the explanatory variable:

$$E(u|x) = E(u) = 0$$

This is the key assumption for showing that OLS is unbiased, with the zero value being of no importance once we assume $E(u|x)$ does not change with x . Note that we can compute the OLS estimates whether or not this assumption holds, or even if there is an underlying population model.²⁹

So, how do we show $\widehat{\beta}_1$ is an unbiased estimate of β_1 (Equation 37)? We need to show that under the four assumptions we just outlined, the expected value of $\widehat{\beta}_1$, when averaged across random samples, will center on β_1 . In other words, unbiasedness has to be understood as related to repeated sampling. We will discuss the answer as a series of steps.

Step 1: Write down a formula for $\widehat{\beta}_1$. It is convenient to use the $\frac{C(x,y)}{V(x)}$ form:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Now get rid of some of this notational clutter by defining $\sum_{i=1}^n (x_i - \bar{x})^2 = SST_x$ (i.e., the total variation in the x_i). Rewrite as:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})y_i}{SST_x}$$

Step 2: Replace each y_i with $y_i = \beta_0 + \beta_1 x_i + u_i$ which uses the first linear assumption and the fact that we have sampled data (our

²⁹ We will focus on $\widehat{\beta}_1$. There are a few approaches to showing unbiasedness. One explicitly computes the expected value of $\widehat{\beta}_1$ conditional on $x, \{x_i : i = 1, \dots, n\}$. Even though this is the more proper way to understand the problem, technically we can obtain the same results by treating the conditioning variables as if they were fixed in repeated samples. That is, to treat the x_i as nonrandom in the derivation. So, the randomness in $\widehat{\beta}_1$ comes through the u_i (equivalently, the y_i). Nevertheless, it is important to remember that x are random variables and that we are taking expectations conditional on knowing them. The approach that we're taking is called sometimes “fixed in repeated samples”, and while not realistic in most cases, it gets us to the same place. We use it as a simplifying device because ultimately this chapter is just meant to help you understand this traditional pedagogy better.

second assumption). The numerator becomes:

$$\begin{aligned}
 \sum_{i=1}^n (x_i - \bar{x}) y_i &= \sum_{i=1}^n (x_i - \bar{x})(\beta_0 + \beta_1 x_i + u_i) \\
 &= \beta_0 \sum_{i=1}^n (x_i - \bar{x}) + \beta_1 \sum_{i=1}^n (x_i - \bar{x}) x_i + \sum_{i=1}^n (x_i - \bar{x}) u_i \\
 &= 0 + \beta_1 \sum_{i=1}^n (x_i - \bar{x})^2 + \sum_{i=1}^n (x_i - \bar{x}) u_i \\
 &= \beta_1 SST_x + \sum_{i=1}^n (x_i - \bar{x}) u_i
 \end{aligned}$$

Note, we used $\sum_{i=1}^n (x_i - \bar{x}) = 0$ and $\sum_{i=1}^n (x_i - \bar{x}) x_i = \sum_{i=1}^n (x_i - \bar{x})^2$ to do this.³⁰

We have shown that:

$$\begin{aligned}
 \hat{\beta}_1 &= \frac{\beta_1 SST_x + \sum_{i=1}^n (x_i - \bar{x}) u_i}{SST_x} \\
 &= \beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x}) u_i}{SST_x}
 \end{aligned}$$

Note how the last piece is the slope coefficient from the OLS regression of u_i on $x_i, i : 1, \dots, n$.³¹ We cannot do this regression because the u_i are not observed. Now define $w_i = \frac{(x_i - \bar{x})}{SST_x}$ so that we have the following:

$$\hat{\beta}_1 = \beta_1 + \sum_{i=1}^n w_i u_i$$

Note the following things that this showed: first, $\hat{\beta}_1$ is a linear function of the unobserved errors, u_i . The w_i are all functions of $\{x_1, \dots, x_n\}$. Second, the random difference between β_1 and the estimate of it, $\hat{\beta}_1$, is due to this linear function of the unobservables.

Step 3: Find $E(\hat{\beta}_1)$. Under the random sampling assumption and the zero conditional mean assumption, $E(u_i | x_1, \dots, x_n) = 0$, that means conditional on each of the x variables:

$$E(w_i u_i | x_1, \dots, x_n) = w_i E(u_i | x_1, \dots, x_n) = 0$$

because w_i is a function of $\{x_1, \dots, x_n\}$. This would be true if in the population u and x are correlated.

Now we can complete the proof: conditional on $\{x_1, \dots, x_n\}$,

³⁰ Told you we would use this result a lot.

³¹ I find it interesting that we see so many $\frac{\text{cov}}{\text{var}}$ terms when working with regression. It shows up constantly. Keep your eyes peeled.

$$\begin{aligned}
E(\widehat{\beta_1}) &= E\left(\beta_1 + \sum_{i=1}^n w_i u_i\right) \\
&= \beta_1 + \sum_{i=1}^n E(w_i u_i) \\
&= \beta_1 + \sum_{i=1}^n w_i E(u_i) \\
&= \beta_1 + 0 \\
&= \beta_1
\end{aligned}$$

Remember, β_1 is the fixed constant in the population. The estimator, $\widehat{\beta_1}$, varies across samples and is the random outcome: before we collect our data, we do not know what $\widehat{\beta_1}$ will be. Under the four aforementioned assumptions, $E(\widehat{\beta_0}) = \beta_0$ and $E(\widehat{\beta_1}) = \beta_1$.

I find it helpful to be concrete when we work through exercises like this. So let's visualize this in Stata. Let's create a Monte Carlo simulation in Stata. We have the following population model:

$$y = 3 + 2x + u \quad (38)$$

where $x \sim \text{Normal}(0, 9)$, $u \sim \text{Normal}(0, 36)$. Also, x and u are independent. The following Monte Carlo simulation will estimate OLS on a sample of data 1,000 times. The true β parameter equals 2. But what will the average $\widehat{\beta}$ equal when we use repeated sampling?

```

clear all
program define ols, rclass
version 14.2
syntax [, obs(integer 1) mu(real 0) sigma(real 1) ]

clear
drop _all
set obs 10000
gen x = 9*rnor()
gen u = 36*rnor()
gen y = 3 + 2*x + u
reg y x
end

simulate beta=_b[x], reps(1000): ols
su
hist beta

```

Table 7 gives us the mean value of $\widehat{\beta_1}$ over the 1,000 repetitions (repeated sampling). While each sample had a different estimate, the

Variable	Obs	Mean	St. Dev.
beta	1,000	2.000737	0.0409954

average for $\widehat{\beta}_1$ was 2.000737, which is close to the true value of 2 (see Equation 38). The standard deviation in this estimator was 0.0409954, which is close to the standard error recorded in the regression itself.³² Thus we see that the estimate is the mean value of the coefficient from repeated sampling, and the standard error is the standard deviation from that repeated estimation. We can see the distribution of these coefficient estimates in Figure 5.

Table 7: Monte Carlo simulation of OLS

³² The standard error I found from running on one sample of data was 0.0393758.

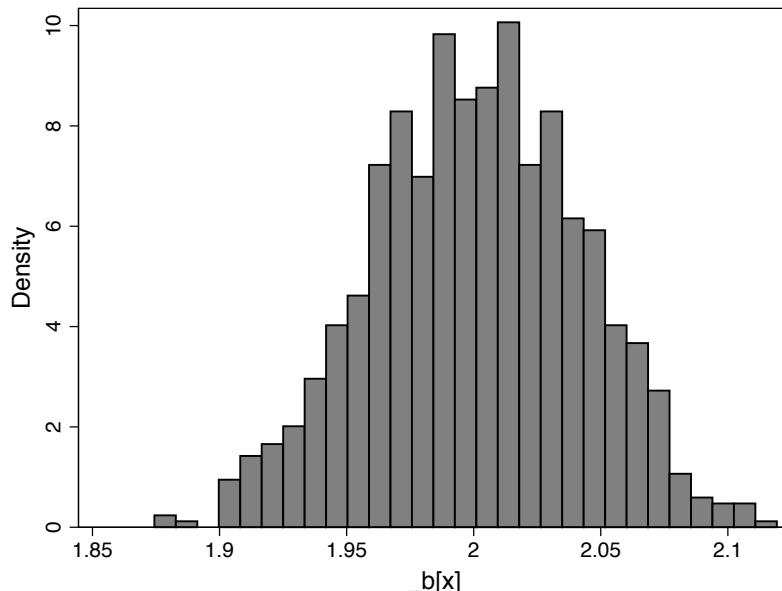


Figure 5: Distribution of coefficients from Monte Carlo simulation.

The problem is, we don't know which kind of sample we have. Do we have one of the "almost exactly 2" samples or do we have one of the "pretty different from 2" samples? We can never know whether we are close to the population value. We hope that our sample is "typical" and produces a slope estimate close to $\widehat{\beta}_1$ but we can't know. Unbiasedness is a property of the procedure of the rule. It is not a property of the estimate itself. For example, say we estimated an 8.2% return on schooling. It is tempting to say 8.2% is an "unbiased estimate" of the return to schooling, but that's incorrect technically. The rule used to get $\widehat{\beta}_1 = 0.082$ is unbiased (if we believe that u is unrelated to schooling) – not the actual estimate itself.

Law of iterated expectations As we said earlier in this chapter, the conditional expectation function (CEF) is the mean of some outcome y with some covariate x held fixed. Now we focus more intently on this function.³³ Let's get the notation and some of the syntax out of the way. As noted earlier, we write the CEF as $E(y_i|x_i)$. Note that the CEF is explicitly a function of x_i . And because x_i is random, the CEF is random – although sometimes we work with particular values for x_i , like $E(y_i|x_i = 8 \text{ years schooling})$ or $E(y_i|x_i = \text{Female})$. When there are treatment variables, then the CEF takes on two values: $E(y_i|d_i = 0)$ and $E(y_i|d_i = 1)$. But these are special cases only.

An important complement to the CEF is the law of iterated expectations (LIE). This law says that an unconditional expectation can be written as the unconditional average of the CEF. In other words $E(y_i) = E\{E(y_i|x_i)\}$. This is a fairly simple idea to grasp. What it states is that if you want to know the unconditional expectation of some random variable y , you can simply calculate the weighted sum of all conditional expectations with respect to some covariate x . Let's look at an example. Let's say that average GPA for females is 3.5, average GPA for males is a 3.2, half the population is females, and half is males. Then:

$$\begin{aligned} E[\text{GPA}] &= E\{E(\text{GPA}_i|\text{Gender}_i)\} \\ &= (0.5 \times 3.5) + (3.2 \times 0.5) \\ &= 3.35 \end{aligned}$$

You probably use LIE all the time and didn't even know it. The proof is not complicated. Let x_i and y_i each be continuously distributed. The joint density is defined as $f_{xy}(u, t)$. The conditional distribution of y given $x = u$ is defined as $f_y(t|x_i = u)$. The marginal densities are $g_y(t)$ and $g_x(u)$.

$$\begin{aligned} E\{E(y|x)\} &= \int E(y|x = u)g_x(u)du \\ &= \int \left[\int tf_{y|x}(t|x = u)dt \right] g_x(u)du \\ &= \int \int tf_{y|x}(t|x = u)g_x(u)dudt \\ &= \int t \left[\int f_{y|x}(t|x = u)g_x(u)du \right] dt \\ &= \int t[f_{x,y}du]dt \\ &= \int tg_y(t)dt \\ &= E(y) \end{aligned}$$

The first line uses the definition of expectation. The second line uses

³³ This section is based heavily on Angrist and Pischke [2009].

the definition of conditional expectation. The third line switches the integration order. The fourth line uses the definition of joint density. The sixth line integrates joint density over the support of x which is equal to the marginal density of y . So restating the law of iterated expectations: $E(y_i) = E\{E(y|x_i)\}$.

CEF Decomposition Property The first property of the CEF we will discuss is the CEF Decomposition Property. The power of LIE comes from the way it breaks a random variable into two pieces – the CEF and a residual with special properties. The CEF Decomposition Property states that

$$y_i = E(y_i|x_i) + \varepsilon_i$$

where (i) ε_i is mean independent of x_i , that is

$$E(\varepsilon_i|x_i) = 0$$

and (ii) ε_i is uncorrelated with any function of x_i .

The theorem says that any random variable y_i can be decomposed into a piece that is “explained by x_i ” (the CEF) and a piece that is left over and orthogonal to any function of x_i . The proof is provided now. I’ll prove the (i) part first. Recall that $\varepsilon_i = y_i - E(y_i|x_i)$ as we will make a substitution in the second line below.

$$\begin{aligned} E(\varepsilon_i|x_i) &= E(y_i - E(y_i|x_i)|x_i) \\ &= E(y_i|x_i) - E(y_i|x_i) \\ &= 0 \end{aligned}$$

The second part of the theorem states that ε_i is uncorrelated with any function of x_i . Let $h(x_i)$ be any function of x_i . Then $E(h(x_i)\varepsilon_i) = E\{h(x_i)E(\varepsilon_i|x_i)\}$ The second term in the interior product is equal to zero by mean independence.³⁴

CEF Prediction Property The second property is the CEF Prediction Property. This states that $E(y_i|x_i) = \arg \min_{m(x_i)} E[(y - m(x_i))^2]$ where $m(x_i)$ is any function of x_i . In words, this states that the CEF is the minimum mean squared error of y_i given x_i . By adding $E(y_i|x_i) - E(y_i|x_i) = 0$ to the right hand side we get

$$[y_i - m(x_i)]^2 = [(y_i - E[y_i|x_i]) + (E(y_i|x_i) - m(x_i))]^2$$

I personally find this easier to follow with simpler notation. So replace this expression with the following terms:

$$(a - b + b - c)^2$$

³⁴ Let’s take a concrete example of this proof. Let $h(x_i) = \alpha + \gamma x_i$. Then take the joint expectation $E(h(x_i)\varepsilon_i) = E[(\alpha + \gamma x_i)\varepsilon_i]$ Then take conditional expectations $E(\alpha|x_i) + E(\gamma|x_i)E(x_i|x_i)E(\varepsilon|x_i) = \alpha + x_i E(\varepsilon|x_i) = 0$ after we pass the conditional expectation through.

Distribute the terms, rearrange, and replace the terms with their original values until you get the following

$$\arg \min (y_i - E(y_i|x_i))^2 + 2(E(y_i|x_i) - m(x_i)) \times (y_i - E(y_i|x_i)) + (E(y_i|x_i) + m(x_i))^2$$

Now minimize the function with respect to $m(x_i)$. When minimizing this function with respect to $m(x_i)$, note that the first term $(y_i - E(y_i|x_i))^2$ doesn't matter because it does not depend on $m(x_i)$. So it will zero out. The second and third terms, though, do depend on $m(x_i)$. So rewrite $2(E(y_i|x_i) - m(x_i))$ as $h(x_i)$. Also set ε_i equal to $[y_i - E(y_i|x_i)]$ and substitute

$$\arg \min \varepsilon_i^2 + h(x_i)\varepsilon_i + [E(y_i|x_i) + m(x_i)]^2$$

Now minimizing this function and setting it equal to zero we get

$$h'(x_i)\varepsilon_i$$

which equals zero by the Decomposition Property.

ANOVA Theory The final property of the CEF that we will discuss is the analysis of variance theorem, or ANOVA. It is simply that the unconditional variance in some random variable is equal to the variance in the conditional expectation plus the expectation of the conditional variance, or

$$V(y_i) = V[E(y_i|x_i)] + E[V(y_i|x_i)]$$

where V is the variance and $V(y_i|x_i)$ is the conditional variance.

Linear CEF Theorem Angrist and Pischke [2009] give several arguments as to why linear regression may be of interest to a practitioner even if the underlying CEF itself is not linear. I will review of those linear theorems now. These are merely arguments to justify the use of linear regression models to approximate the CEF.³⁵

The Linear CEF Theorem is the most obvious theorem of the three that Angrist and Pischke [2009] discuss. Suppose that the CEF itself is linear. Then the population regression is equal to the CEF. This simply states that you should use the population regression to estimate the CEF when you know that the CEF is linear. The proof is provided. If $E(y_i|x_i)$ is linear, then $E(y_i|x_i) = x'\hat{\beta}$ for some K vector $\hat{\beta}$. By the Decomposition Property

$$E(x(y - E(y|x))) = E(x(y - x'\hat{\beta})) = 0$$

Solve this and get $\hat{\beta} = \beta$. Hence $E(y|x) = x'\beta$.

³⁵ Note, Angrist and Pischke [2009] make their arguments for using regression, not based on unbiasedness and the four assumptions that we discussed, but rather because regression approximates the CEF. I want to emphasize that this is a subtly different direction. I included the discussion of unbiasedness, though, to be exhaustive. Just note, there is a slight change in pedagogy though.

Best Linear Predictor Theorem Recall that the CEF is the minimum mean squared error predictor of y given x in the class of all functions according to the CEF prediction property. Given this, the population regression function, $E(X'Y)E(X'X)^{-1}$ is the best that we can do in the class of all linear functions.³⁶ Proof: β solves the population minimum mean squared error problem.

Regression CEF Theorem The function $X\beta$ provides the minimum mean squared error linear approximation to the CEF. That is,

$$\beta = \arg \min_b E\{[E(y_i|x_i) - x'_i b]^2\}$$

Regression anatomy theorem In addition to our discussion of the CEF and regression theorems, we now dissect the regression itself. Here we discuss the **regression anatomy theorem**. The regression anatomy theorem is based on earlier work by [Frisch and Waugh \[1933\]](#) and [Lovell \[1963\]](#).³⁷ I find it more intuitive when thinking through a specific example and offering up some data visualization. In my opinion, the theorem helps us interpret the individual coefficients of a multiple linear regression model. Say that we are interested in the causal effect of family size on labor supply. We want to regress labor supply onto family size:

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

where Y is labor supply and X is family size.

If family size is truly random, then the number of kids is uncorrelated with the unobserved error term. This implies that when we regress labor supply onto family size, our estimate $\hat{\beta}_1$ can be interpreted as the causal effect of family size on labor supply. Visually, we could just plot the regression coefficient in a scatter plot showing all i pairs of data, and the slope coefficient would be the best fit of this data through this data cloud. That slope would tell us the average causal effect of family size on labor supply.

But how do we interpret $\hat{\beta}_1$ if the family size is *not random*? After all, we know from living on planet Earth and having even half a brain that a person's family size is usually chosen, not randomly assigned to them. And oftentimes, it's chosen according to something akin to an optimal stopping rule. People pick both the number of kids to have, as well as when to have them, and in some instance, even attempt to pick the gender, and this is all based on a variety of observed and unobserved economic factors that are directly correlated with the decision to supply labor. In other words, using the language we've been using up til now, it's unlikely that $E(u|X) = E(u) = 0$.

³⁶ Note that $E(X'Y)E(X'X)^{-1}$ is the matrix notation expression of the population regression, or what we have discussed as $\frac{C(X,Y)}{V(X)}$.

³⁷ A helpful proof of the Frisch-Waugh-Lovell theorem can be found at [Lovell \[2008\]](#).

But let's say that we have reason to think that the number of kids is *conditionally* random. That is, for a given person of a certain race and age, any remaining variation in family size across a population is random.³⁸ Then we have the following population model:

$$Y_i = \beta_0 + \beta_1 X_i + \gamma_1 R_i + \gamma_2 A_i + u_i$$

where Y is labor supply, X is family size, R is race, A is age, and u is the population error term.

If we want to estimate the average causal effect of family size on labor supply, then we need two things. First, we need a sample of *data* containing all four of these variables. Without all four of the variables, we cannot estimate this regression model. And secondly, we need for number of kids, X , to be randomly assigned for a given set of race/age.

Now how do we interpret $\hat{\beta}_1$? And for those who like pictures, how might we visualize this coefficient given there's six dimensions to the data? The regression anatomy theorem tells us both what this coefficient estimate actually means, and it also lets us visualize the data in only two dimensions.

To explain the intuition of the regression anatomy theorem, let's write down a population model with multiple variables. Assume that your main multiple regression model of interest is

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \cdots + \beta_K x_{Ki} + e_i \quad (39)$$

Now assume an *auxiliary* regression in which the variable x_{1i} is regressed on all the remaining independent variables

$$x_{1i} = \gamma_0 + \gamma_{k-1} x_{k-1i} + \gamma_{k+1} x_{k+1i} + \cdots + \gamma_K x_{Ki} + f_i \quad (40)$$

and $\tilde{x}_{1i} = x_{1i} - \hat{x}_{1i}$ being the residual from that auxiliary regression. Then the parameter β_1 can be rewritten as:

$$\beta_1 = \frac{C(y_i, \tilde{x}_i)}{V(\tilde{x}_i)} \quad (41)$$

Notice that again we see the coefficient estimate being a scaled covariance, only here the covariance is with respect to the outcome and residual from the auxiliary regression and the scale is the variance of that same residual.

To prove the theorem, note that $E[\tilde{x}_{ki}] = E[x_{ki}] - E[\hat{x}_{ki}] = E[f_i]$, and plug y_i and residual \tilde{x}_{ki} from x_{ki} auxiliary regression into the covariance $cov(y_i, x_{ki})$.

$$\begin{aligned} \beta_k &= \frac{cov(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \cdots + \beta_K x_{Ki} + e_i, \tilde{x}_{ki})}{var(\tilde{x}_{ki})} \\ &= \frac{cov(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \cdots + \beta_K x_{Ki} + e_i, f_i)}{var(f_i)} \end{aligned}$$

³⁸ Almost certainly not a credible assumption, but stick with me.

Since by construction $E[f_i] = 0$, it follows that the term $\beta_0 E[f_i] = 0$. Since f_i is a linear combination of all the independent variables with the exception of x_{ki} , it must be that

$$\beta_1 E[f_i x_{1i}] = \dots = \beta_{k-1} E[f_i x_{k-1i}] = \beta_{k+1} E[f_i x_{k+1i}] = \dots = \beta_K E[f_i x_{Ki}] = 0$$

Consider now the term $E[e_i f_i]$. This can be written as

$$\begin{aligned} E[e_i f_i] &= E[e_i f_i] \\ &= E[e_i \tilde{x}_{ki}] \\ &= E[e_i(x_{ki} - \hat{x}_{ki})] \\ &= E[e_i x_{ki}] - E[e_i \tilde{x}_{ki}] \end{aligned}$$

Since e_i is uncorrelated with any independent variable, it is also uncorrelated with x_{ki} . Accordingly, we have $E[e_i x_{ki}] = 0$. With regard to the second term of the subtraction, substituting the predicted value from the x_{ki} auxiliary regression, we get

$$E[e_i \tilde{x}_{ki}] = E[e_i(\hat{\gamma}_0 + \hat{\gamma}_1 x_{1i} + \dots + \hat{\gamma}_{k-1} x_{k-1i} + \hat{\gamma}_{k+1} x_{k+1i} + \dots + \hat{\gamma}_K x_{Ki})]$$

Once again, since e_i is uncorrelated with any independent variable, the expected value of the terms is equal to zero. Then it follows that $E[e_i f_i] = 0$.

The only remaining term then is $[\beta_k x_{ki} f_i]$ which equals $E[\beta_k x_{ki} \tilde{x}_{ki}]$ since $f_i = \tilde{x}_{ki}$. The term x_{ki} can be substituted using a rewriting of the auxiliary regression model, x_{ki} , such that

$$x_{ki} = E[x_{ki}|X_{-k}] + \tilde{x}_{ki}$$

This gives

$$\begin{aligned} E[\beta_k x_{ki} \tilde{x}_{ki}] &= \beta_k E[\tilde{x}_{ki}(E[x_{ki}|X_{-k}] + \tilde{x}_{ki})] \\ &= \beta_k \{E[\tilde{x}_{ki}^2] + E[(E[x_{ki}|X_{-k}] \tilde{x}_{ki})]\} \\ &= \beta_k \text{var}(\tilde{x}_{ki}) \end{aligned}$$

which follows directly from the orthogonality between $E[x_{ki}|X_{-k}]$ and \tilde{x}_{ki} . From previous derivations we finally get

$$\text{cov}(y_i, \tilde{x}_{ki}) = \beta_k \text{var}(\tilde{x}_{ki})$$

which completes the proof.

I find it helpful to visualize things. Let's look at an example in Stata.

```
. ssc install reganat, replace
. sysuse auto.dta, replace
. regress price length
. regress price length weight headroom mpg
. reganat price length weight headroom mpg, dis(length) biline
```

Let's walk through the regression output. The first regression of

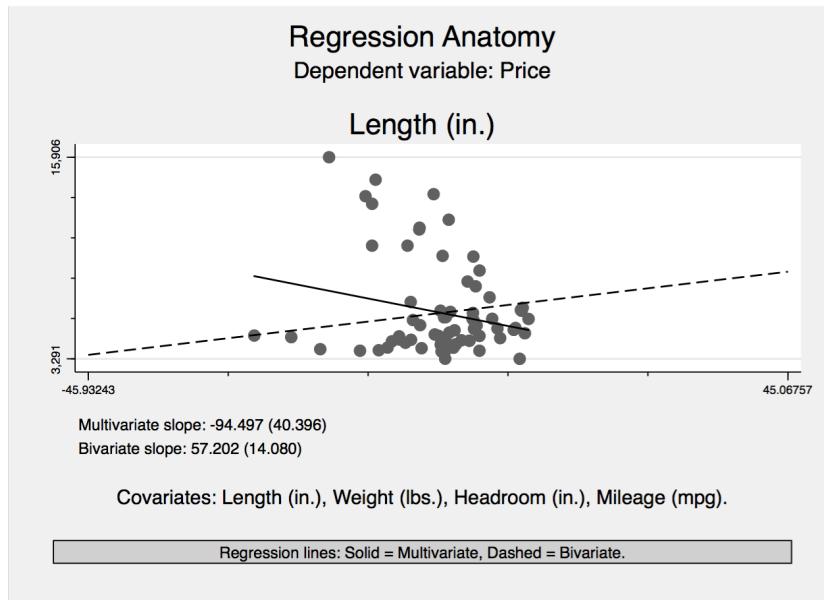


Figure 6: Regression anatomy display.

price on length yields a coefficient of 57.20 on length. But notice the output from the fourth line. The effect on length is -94.5 . The first regression is a bivariate regression and gives a positive slope, but the second regression is a multivariate regression and yields a negative slope.

One of the things we can do with regression anatomy (though this isn't its main purpose) is visualize this negative slope from the multivariate regression in nevertheless two dimensional space. Now how do we visualize this first multivariate slope coefficient, given our data has four dimensions? We run the auxiliary regression, use the residuals, and then calculate the slope coefficient as $\frac{\text{cov}(y_i, \hat{x}_i)}{\text{var}(\hat{x}_i)}$. We can also show scatter plots of these auxiliary residuals paired with their outcome observations and slice the slope through them (Figure 6). Notice that this is a useful way to preview the multidimensional correlation between two variables from a multivariate regression.

And as we discussed before, the solid black line is negative while the slope from the bivariate regression is positive. The regression anatomy theorem shows that these two estimators – one being a multivariate OLS and the other being a bivariate regression price and a residual – are identical.

Variance of the OLS Estimators In this chapter we discuss inference under a variety of situations. Under the four assumptions we mentioned earlier, the OLS estimators are unbiased. But these assumptions are not sufficient to tell us anything about the variance in the estimator itself. These assumptions help inform our beliefs that the estimated coefficients, on average, equal the parameter values themselves. But to speak intelligently about the variance of the estimator, we need a measure of dispersion, or spread, in the sampling distribution of the estimators. As we've been saying, this leads us to the variance and ultimately the standard deviation. We could characterize the variance of the OLS estimators under the four assumptions. But for now, it's easiest to introduce an assumption that simplifies the calculations. We'll keep the assumption ordering we've been using and call this the fifth assumption.

The fifth assumption is the *homoskedasticity* or constant variance assumption. This assumption stipulates that our population error term, u , has the same variance given any value of the explanatory variable, x . Formally, it's:

$$V(u|x) = \sigma^2 > 0 \quad (42)$$

where σ is some finite, positive number. Because we assume the zero conditional mean assumption, whenever we assume homoskedasticity, we can also write:

$$E(u^2|x) = \sigma^2 = E(u^2) \quad (43)$$

Now, under the first, fourth and fifth assumptions, we can write:

$$\begin{aligned} E(y|x) &= \beta_0 + \beta_1 x \\ V(y|x) &= \sigma^2 \end{aligned} \quad (44)$$

So the average, or expected, value of y is allowed to change with x , but the variance does not change with x . The constant variance assumption may not be realistic; it must be determined on a case-by-case basis.

Theorem: Sampling variance of OLS. Under assumptions 1 and 2,

we get:

$$\begin{aligned} V(\widehat{\beta}_1|x) &= \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \frac{\sigma^2}{SST_x} \end{aligned} \quad (45)$$

$$V(\widehat{\beta}_0|x) = \frac{\sigma^2 (\frac{1}{n} \sum_{i=1}^n x_i^2)}{SST_x} \quad (46)$$

To show this, write, as before,

$$\widehat{\beta}_1 = \beta_1 + \sum_{i=1}^n w_i u_i \quad (47)$$

where $w_i = \frac{(x_i - \bar{x})}{SST_x}$. We are treating this as nonrandom in the derivation. Because β_1 is a constant, it does not affect $V(\widehat{\beta}_1)$. Now, we need to use the fact that, for uncorrelated random variables, the variance of the sum is the sum of the variances. The $\{u_i : i = 1, \dots, n\}$ are actually independent across i and are uncorrelated. Remember: if we know x , we know w . So:

$$V(\widehat{\beta}_1|x) = Var(\beta_1 + \sum_{i=1}^n w_i u_i|x) \quad (48)$$

$$= Var\left(\sum_{i=1}^n w_i u_i|x\right) \quad (49)$$

$$= \sum_{i=1}^n Var(w_i u_i|x) \quad (50)$$

$$= \sum_{i=1}^n w_i^2 Var(u_i|x) \quad (51)$$

$$= \sum_{i=1}^n w_i^2 \sigma^2 \quad (52)$$

$$= \sigma^2 \sum_{i=1}^n w_i^2 \quad (53)$$

where the penultimate equality condition used the fifth assumption so that the variance of u_i does not depend on x_i . Now we have:

$$\sum_{i=1}^n w_i^2 = \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{SST_x^2} \quad (54)$$

$$= \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{SST_x^2} \quad (55)$$

$$= \frac{SST_x}{SST_x^2} \quad (56)$$

$$= \frac{1}{SST_x} \quad (57)$$

We have shown:

$$V(\widehat{\beta}_1) = \frac{\sigma^2}{SST_x} \quad (58)$$

A couple of points. First, this is the “standard” formula for the variance of the OLS slope estimator. It is *not* valid if the fifth assumption (“homoskedastic errors”) doesn’t hold. The homoskedasticity assumption is needed, in other words, to derive this standard formula. But, the homoskedasticity assumption is *not* used to show unbiasedness of the OLS estimators. That requires only the first four assumptions we discussed.

Usually, we are interested in β_1 . We can easily study the two factors that affect its variance: the numerator and the denominator.

$$V(\widehat{\beta}_1) = \frac{\sigma^2}{SST_x} \quad (59)$$

As the error variance increases – that is, as σ^2 increases – so does the variance in our estimator. The more “noise” in the relationship between y and x (i.e., the larger the variability in u) – the harder it is to learn something about β_1 . By contrast, more variation in $\{x_i\}$ is a *good* thing. As $SST_x \uparrow$, $V(\widehat{\beta}_1) \downarrow$.

Notice that $\frac{SST_x}{n}$ is the sample variance in x . We can think of this as getting close to the population variance of x , σ_x^2 , as n gets large. This means:

$$SST_x \approx n\sigma_x^2 \quad (60)$$

which means that as n grows, $V(\widehat{\beta}_1)$ shrinks at the rate of $\frac{1}{n}$. This is why more data is a good thing – because it shrinks the sampling variance of our estimators.

The standard deviation of $\widehat{\beta}_1$ is the square root of the variance. So:

$$sd(\widehat{\beta}_1) = \frac{\sigma}{\sqrt{SST_x}} \quad (61)$$

This turns out to be the measure of variation that appears in confidence intervals and test statistics.

Next we look at estimating the error variance. In the formula, $V(\widehat{\beta}_1) = \frac{\sigma^2}{SST_x}$, we can compute SST_x from $\{x_i : i = 1, \dots, n\}$. But we need to estimate σ^2 . Recall that $\sigma^2 = E(u^2)$. Therefore, if we could observe a sample on the errors, $\{u_i : i = 1, \dots, n\}$, an unbiased estimator of σ^2 would be the sample average:

$$\frac{1}{n} \sum_{i=1}^n u_i^2 \quad (62)$$

But this isn’t an estimator that we can compute from the data we observe because u_i are unobserved. How about replacing each u_i

with its “estimate”, the OLS residual \hat{u}_i :

$$u_i = y_i - \beta_0 - \beta_1 x_i \quad (63)$$

$$\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \quad (64)$$

Whereas u_i cannot be computed, \hat{u}_i can be computed from the data because it depends on the estimators, $\hat{\beta}_0$ and $\hat{\beta}_1$. But, except by fluke, $u_i \neq \hat{u}_i$ for any i .

$$\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \quad (65)$$

$$= (\beta_0 + \beta_1 x_i + u_i) - \hat{\beta}_0 - \hat{\beta}_1 x_i \quad (66)$$

$$= u_i - (\hat{\beta}_0 - \beta_0) - (\hat{\beta}_1 - \beta_1) x_i \quad (67)$$

Note that $E(\hat{\beta}_0) = \beta_0$ and $E(\hat{\beta}_1) = \beta_1$, but the estimators almost always differ from the population values in a sample. So what about this as an estimator of σ^2 ?

$$\frac{1}{n} \sum_{i=1}^n \hat{u}_i^2 = \frac{1}{n} SSR \quad (68)$$

It is a true estimator and easily computed from the data after OLS.

As it turns out, this estimator is slightly biased: its expected value is a little less than σ^2 . The estimator does not account for the two restrictions on the residuals used to obtain $\hat{\beta}_0$ and $\hat{\beta}_1$:

$$\sum_{i=1}^n \hat{u}_i = 0 \quad (69)$$

$$\sum_{i=1}^n x_i \hat{u}_i = 0 \quad (70)$$

There is no such restriction on the unobserved errors. The unbiased estimator, therefore, of σ^2 uses a **degrees of freedom** adjustment. The residuals have only $n - 2$ degrees-of-freedom, not n . Therefore:

$$\hat{\sigma}^2 = \frac{1}{n-2} SSR \quad (71)$$

We now propose the following theorem. **The Unbiased Estimator of σ^2** under the first five assumptions is:

$$E(\hat{\sigma}^2) = \sigma^2 \quad (72)$$

In regression output, this is the usually reported:

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} \quad (73)$$

$$= \sqrt{\frac{SSR}{(n-2)}} \quad (74)$$

This is an estimator of $sd(u)$, the standard deviation of the population error. One small glitch is that $\hat{\sigma}$ is not unbiased for σ .³⁹ This will

³⁹ There does exist an unbiased estimator of σ but it's tedious and hardly anyone in economics seems to use it. See Holtzman [1950].

not matter for our purposes. $\hat{\sigma}$ is called the **standard error of the regression**, which means it is an estimate of the standard deviation of the error in the regression. Stata calls it the **root mean squared error**.

Given $\hat{\sigma}$, we can now estimate $sd(\hat{\beta}_1)$ and $sd(\hat{\beta}_0)$. The estimates of these are called the **standard errors** of the $\hat{\beta}_j$. We will use these a lot. Almost all regression packages report the standard errors in a column next to the coefficient estimates. We just plug $\hat{\sigma}$ in for σ :

$$se(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{SST_x}} \quad (75)$$

where both the numerator and denominator are computed from the data. For reasons we will see, it is useful to report the standard errors below the corresponding coefficient, usually in parentheses.

Cluster robust standard errors Some phenomena do not affect observations individually, but rather, affect groups of observations which contain individuals. And then it affects those individuals within the group in a common way. Say you wanted to estimate the effect of class size on student achievement, but you know that there exist unobservable things (like the teacher) which affects all the students equally. If we can commit to independence of these unobservables across classes, but individual student unobservables are correlated within a class, then we have a situation where we need to cluster the standard errors. Here's an example:

$$y_{ig} = x'_{ig}\beta + \varepsilon_{ig} \text{ where } 1, \dots, G$$

and

$$E[\varepsilon_{ig}\varepsilon'_{jg}]$$

which equals zero if $g = g'$ and equals $\sigma_{(ij)g}$ if $g \neq g'$.

Let's stack the data by cluster first.

$$y_g = x'_g\beta + \varepsilon_g$$

The OLS estimator is still $\hat{\beta} = E[X'X]^{-1}X'Y$. We just stacked the data which doesn't affect the estimator itself. But it does change the variance.

$$V(\beta) = E[[X'X]^{-1}X'\Omega X[X'X]^{-1}]$$

With this in mind, we can now write the variance-covariance matrix for clustered data as

$$\hat{V}(\hat{\beta}) = [X'X]^{-1} \left[\sum_{i=1}^G x'_g \hat{\varepsilon}_g \hat{\varepsilon}'_g \right] [X'X]^{-1}$$

Directed acyclical graphs

Here we take a bit of a detour, because this material is not commonly featured in the economist's toolbox. It is nonetheless extremely valuable, and worth spending some time learning because I will try to convince you that these graphical models can help you to identify causal effects in observational data.

The history of graphical causal modeling in science goes back to Phillip Wright, an economist and the father of Sewell Wright, the father of modern genetics. Sewell developed path diagrams for genetics and Philip, we believe, adapted them for econometric identification [Matsueda, 2012].⁴⁰

The use of graphs in causal modeling has been largely ignored by the economics profession with only a few exceptions [Heckman and Pinto, 2015]. It was not revitalized for the purposes of causal inference until Judea Pearl began developing his own unique theory of causation using them [Pearl, 2009]. Pearl's influence has been immense outside of economics, including many of the social sciences, but few economists are familiar with him or use graphical models in their work. Since I think graphical models are immensely helpful for designing a credible identification strategy, I have chosen to include these models for your consideration. We will now have a simple review of graphical models, one of Pearl's contributions to the theory of causal inference.⁴¹

⁴⁰ We will discuss Wright again in the chapter on instrumental variables.

⁴¹ This section is heavily influenced by Morgan and Winship [2014].

Introduction to DAG notation

Before we begin, I'd like to discuss some limitations to the directed acyclical graphical (DAG) representation of causality. The first to note in the DAG notation is causality runs in one direction. There are no cycles in a DAG. To show reverse causality, one would need to create multiple nodes, most likely with two versions of the same node separated by a time index. Secondly, DAGs may not be built to handle simultaneity according to Heckman and Pinto [2015]. But with those limitations in mind, we proceed forward as I have found

DAGs to be extremely valuable otherwise.

A DAG is a way of modeling a causal effect using graphs. The DAG represents these causal effects through a set of nodes and arrows, or directed edges. For a complete understanding of the notation, see Pearl [2009]. I will use a modified shorthand that I believe is sufficient for my purposes in the book.

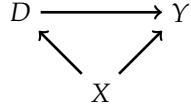
A DAG contains nodes which represent random variables. These random variables are assumed to be created by some data generating process that is often left out of the DAG itself, though not always. I leave them out because it clutters the graph unnecessarily. Arrows represent a causal effect between two random variables moving in the intuitive direction of the arrow. The direction of the arrow captures cause and effect, in other words. Causal effects can happen in two ways. They can either be direct (e.g., $D \rightarrow Y$), or they can be mediated by a third variable (e.g., $D \rightarrow X \rightarrow Y$). When they are mediated by the third variable, technically speaking we are not capturing the effect of D on Y , but rather we are capturing a sequence of events originating with D , which may or may not be important to you depending on the question you're asking.

A DAG is meant to be a complete description of all causal relationships relevant to some phenomena relevant to the effect of D on Y . What makes the DAG distinctive is both the explicit commitment to a causal effect pathway, but also the complete commitment to the *lack* of a causal pathway represented by missing arrows. A complete DAG will have all direct causal effects among the variables in the graph, as well as all common causes of any pair of variables in the graph.

At this point, you may be wondering, “where does the DAG come from?” It’s an excellent question. A DAG is a theoretical representation of some phenomena, and it comes from a variety of sources. Examples would include economic theory, economic models, your own observations and experiences, literature reviews, as well as your own intuition and hypotheses.

I will argue that the DAG, at minimum, is useful for a few reasons. One, it is helpful for students to better understand research designs and estimators for the first time. This is, in my experience, especially true for instrumental variables which has a very intuitive DAG representation. Two, through concepts such as the backdoor criterion and collider bias, a well-designed DAG can help you develop a credible research design for identifying the causal effects of some intervention.

A Simple DAG Let’s begin with a concrete example. Consider the following DAG. We begin with a basic DAG to illustrate a few ideas, but will expand it to slightly more complex ones later.

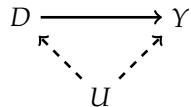


In this DAG, we have three random variables: X , D and Y . There is a direct *path* from D to Y , which represents a causal effect. There are two paths from D to Y – one direct path, and one *backdoor path*. The direct path, or causal effect, is $D \rightarrow Y$.

The idea of the backdoor path is one of the most important things that we learn from the DAG. It is similar to the notion of omitted variable bias in that it represents a determinant of some outcome that is itself correlated with a variable of interest. Just as not controlling for a variable like that in a regression creates omitted variable bias, leaving a backdoor open creates bias. The backdoor path is $D \leftarrow X \rightarrow Y$. We therefore call X a *confounder* in the sense that because it jointly determines D and Y , it confounds our ability to discern the effect of D on Y in naive comparisons.

Think of the backdoor path like this: sometimes when D takes on different values, Y takes on different values because D causes Y . But sometimes D and Y take on different values because X takes on different values, and that bit of the correlation between D and Y is purely spurious. The existence of two causal pathways is contained within correlation between D and Y . When a backdoor path has a confounder on it and no “collider”, we say that backdoor path is *open*.⁴²

Let's look at a second DAG, this one more problematic than the one before. In the previous example, X was observed. We know it was observed because the direct edges from X to D and Y were solid lines. But sometimes there exists a confounder that is unobserved, and when there is, we represent its direct edges with dashed lines. Consider the following DAG:

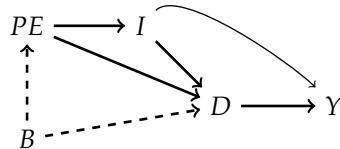


Same as before, U is a noncollider along the backdoor path from D to Y , but unlike before, U is unobserved to the researcher. It exists, but it may simply be missing from the dataset. In this situation, there are two pathways from D to Y . There's the direct pathway, $D \rightarrow Y$, which is the causal effect, and there's the backdoor pathway $D \leftarrow U \rightarrow Y$. And since U is unobserved, that backdoor pathway is *open*.

Let's now move to another example, one that is slightly more realistic. A traditional in labor economics is whether college education

⁴² More on colliders in a moment.

increases earnings. According to the Becker human capital model [Becker, 1994], education increases one's marginal product, and since workers are paid their marginal product in competitive markets, it also increases their earnings. But, college education is not random; it is optimally chosen given subjective preferences and resource constraints. We will represent that with the following DAG. As always, let D be the treatment (e.g., college education) and Y be the outcome of interest (e.g., earnings). Furthermore, let PE be parental education, I be family income, and B be unobserved background factors, such as genetics, family environment, mental ability, etc.



This DAG is telling a story. Can you interpret that story for yourself?

Here is my interpretation. Each person has some background. It's not contained in the most datasets, as it measures things like intelligence, contentiousness, mood stability, motivation, family dynamics, and other environmental factors. Those environmental factors are likely correlated between parent and child, and therefore are subsumed in the variable B . Background causes a relevant parent to herself choose some level of education, and that choice also causes the child to choose a level of education through a variety of channels. First, there is the shared background factors, B . Those background factors cause the child to herself choose a level of education, just as it had with the parent. Second, there's a direct effect, perhaps through simple modeling of achievement, a kind of peer effect. And third, there's the effect that parental education has on family earnings, I , which in turn affects how much schooling the child receives. Family earnings may itself affect earnings through bequests and other transfers, as well as external investments in the child's productivity.

This is a simple story to tell, and the DAG tells it well, but I want to alert your attention to some subtle points contained in this DAG. One, notice that B has no direct effect on the child's earnings except through its effect on schooling. Is this realistic, though? Economists have long maintained that unobserved ability both determines how much schooling a child gets, but also directly affects their earnings, insofar as intelligence and motivation can influence careers. But in this DAG, there is no relationship between background and earnings, which is itself an *assumption*.

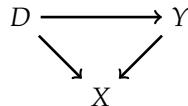
Now that we have a DAG, what do we do? We want to list out all the direct paths and indirect paths (i.e., backdoor paths) between D

and Y .

1. $D \rightarrow Y$ (the causal effect of education on earnings)
2. $D \leftarrow I \rightarrow Y$ (backdoor path # 1)
3. $D \leftarrow PE \rightarrow I \rightarrow Y$ (backdoor path # 2)
4. $D \leftarrow B \rightarrow PE \rightarrow I \rightarrow Y$ (backdoor path # 3)

Thus, we have four paths between D and Y : one direct causal effect and three backdoor paths. And since none of the variables along the backdoor paths are *colliders*, each of these backdoors paths are *open*, creating systematic and independent correlations between D and Y .

Colliding But what is this term “collider”. It’s an unusual term, one you may have never seen before, so let’s introduce it with another example. We’ll use a simple DAG to illustrate what a collider is.



Notice in this graph there are two paths from D to Y as before. There’s the direct (causal) path, $D \rightarrow Y$. And there’s the backdoor path, $D \rightarrow X \leftarrow Y$. Notice the subtle difference in this backdoor path than in the previous one. This time the X has two arrows from D and Y point to it. X on this backdoor path is called a “collider” (as opposed to a confounder) because D and Y ’s causal effects are *colliding* at X . But first, let’s list all paths from D to Y .

1. $D \rightarrow Y$ (causal effect of D on Y)
2. $D \rightarrow X \leftarrow Y$ (backdoor path # 1)

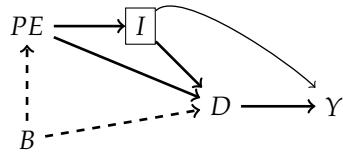
Here we have one backdoor path. And because along that backdoor path is a collider, it is currently *closed*. Colliders, when they are left alone, always close a specific backdoor path.

Backdoor criterion Open backdoor paths create systematic, non-causal correlations between D and Y . Thus, usually our goal is to close that specific backdoor path. And if we can close all backdoor paths, then we can isolate the causal effect of D on Y using one of the research designs and identification strategies discussed in this book. So how do we close a backdoor path?

There are two ways to close a backdoor path. First, if you have a confounder that has created an open backdoor path, then you can close that path by *conditioning* on the confounder. Conditioning

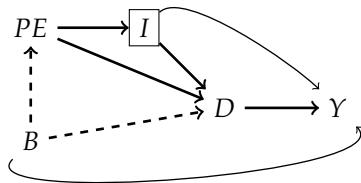
requires holding the variable fixed using something like subclassification, matching, regression, or some other method. It is equivalent to “controlling for” the variable in a regression. The second way to close a backdoor path is if along that backdoor path appears a collider. Since colliders always close backdoor paths, and conditioning on a collider always opens a backdoor path, you want to leave colliders alone. That is, don’t control for colliders in any way, and you will have closed that backdoor path.

When all backdoor paths have been closed, we say that you have met the *backdoor criterion* through some conditioning strategy. Let’s formalize it: a set of variables X satisfies the backdoor criterion in a DAG if and only if X blocks every path between confounders that contain an arrow from D to Y . Let’s review our original DAG involving parental education, background and earnings.



The minimally sufficient conditioning strategy necessary to achieve the backdoor criterion is the control for I , because I appeared as a non-collider along every backdoor path (see earlier).

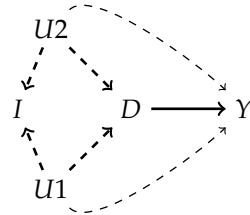
But maybe in hearing this story, and studying it for yourself by reviewing the literature and the economic theory surrounding it, you are skeptical of this DAG. Specifically, you are skeptical that B has no relationship to Y except through D or PE . That skepticism leads you to believe that there should be a *direct* connection from B to Y , not merely one mediated through own education.



Note that including this new backdoor path has created a problem because no longer is our conditioning strategy satisfying the backdoor criterion. Even controlling for I , there still exists spurious correlations between D and Y , and without more information about the nature of $B \rightarrow Y$ and $B \rightarrow D$, we cannot say much more about the partial correlation between D and Y – only that it’s biased.

In our earlier DAG with collider bias, we conditioned on some variable X that was a collider – specifically, though, it was a descen-

dent of D and Y . But sometimes, colliders are more subtle. Let's consider the following scenario. Again, let D and Y be child schooling and child earnings. But this time we introduce three new variables – U_1 , which is father's unobserved genetic ability, U_2 , which is mother's unobserved genetic ability, and I which is joint family income. Assume that I is observed, but U_i is unobserved for both parents.



Notice in this DAG, there are several backdoor paths from D to Y . They are:

1. $D \leftarrow U_2 \rightarrow Y$
2. $D \leftarrow U_1 \rightarrow Y$
3. $D \leftarrow U_1 \rightarrow I \leftarrow U_2 \rightarrow Y$
4. $D \leftarrow U_2 \rightarrow I \leftarrow U_1 \rightarrow Y$

Notice, the first two are open backdoor paths, and as such, cannot be closed because U_1 and U_2 are not observed. But what if we controlled for I anyway? Controlling for I only makes matters worse, because then it opens the third and fourth backdoor paths, as I was a collider along both of them. It does not appear that *any* conditioning strategy could meet the backdoor criterion in this DAG.

So to summarize, satisfying the backdoor criterion requires simply a few steps. First, write down all paths – both directed and backdoor paths – between D and Y . Second, note whether each backdoor path is open or closed by checking for whether there are any colliders along those backdoor paths or confounders. Third, check whether you can close all backdoor paths through some conditioning strategy. If you can do that, then that conditioning strategy satisfies the backdoor criterion and thus you can identify the causal effect of D on Y .

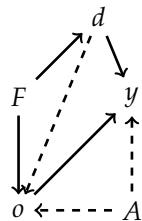
Examples of collider bias: Gender disparities controlling for occupation

The issue of conditioning on a collider is important, so how do we know if we have that problem or not? No dataset is going to come with a flag saying "collider" and "confounder". Rather, the only

way to know if you have satisfied the backdoor criterion is with a DAG, and a DAG requires a model. It requires in-depth knowledge of the data generating process for the variables in your DAG, but it also requires ruling out pathways too. And the only way to rule out pathways is through logic and models. There is no way to avoid it – all empirical work requires theory to guide the work. Otherwise, how do you know if you've conditioned on a collider or a noncollider? Put differently, you cannot identify treatment effects without making assumptions.

Collider bias is a difficult concept to understand at first, so I've included a couple of examples to help you sort through it. So let's first examine a real world example. It is common to hear someone deny the existence of gender disparities in earnings by saying that once occupation or other characteristics of a job are conditioned on, the wage disparity disappears or gets smaller. For instance, the NYT claimed that Google systematically underpaid its female employees. But Google responded that their data showed that when you take "location, tenure, job role, level and performance" into consideration, female pay is basically identical to that of male counterparts. In other words, controlling for characteristics of the job, women received the same pay.

But what if one of the ways in which gender discrimination creates gender disparities in earnings is *through* occupational sorting? Then naive regressions of wages onto a gender dummy controlling for occupation characteristics will be biased towards zero, thus understating the degree of discrimination in the marketplace. Put differently, when there exists occupational sorting based on unobserved ability then assuming gender discrimination we cannot identify the actual discrimination effect controlling for occupation. Let's first give a DAG to illustrate the problem.



Notice that there is in fact no effect of females on earnings, because they are assumed to be just as productive of males. Thus if we could control for discrimination, we'd get a coefficient of zero as in this example women are just as productive as men.

But in this example, we aren't interested in estimating the effect of female on earnings; we are interested in estimating the effect

of discrimination itself. Now you can see several backdoor paths between discrimination and earnings. They are:

1. $d \leftarrow F \rightarrow o \rightarrow y$
2. $d \rightarrow o \rightarrow y$
3. $d \leftarrow F \rightarrow o \leftarrow A$
4. $d \rightarrow o \leftarrow A \rightarrow y$

So let's say we regress y onto d (which will always pick up the discrimination effect). This is biased because it picks up the effect of discrimination on occupation and earnings, as well as gender's effect on occupation and earnings. So naturally, we might want to control for occupation, but notice when we do this, we close down those two backdoor paths *but open* a new path (the last one). That is because $F \rightarrow o \leftarrow A \rightarrow y$ has a collider (o). So when we control for occupation, we open up a new path. This is the reason we cannot merely control for occupation. Such a control ironically introduces new patterns of bias.

What is needed rather is to control for occupation *and* ability, but since ability is unobserved, we cannot do that, and therefore we do not possess an identification strategy that satisfies the backdoor criterion. Let's now look at Stata code created by Erin Hengel at the University of Liverpool which she has graciously lent to me with permission to reproduce here.⁴³

* Create confounding bias for female occupation and gender gap
clear all
set obs 10000

* Half of the population is female.

generate female = runiform()>=0.5

* Innate ability is independent of gender.

generate ability = rnormal()

* All women experience discrimination.

generate discrimination = female

* Continuum of occupations ranked monotonically according to ability, conditional

* on discrimination—i.e., higher ability people are allocated to higher ranked

⁴³ Erin has done very good work on gender discrimination. See her website for more of this <http://www.erinhengel.com>.

* occupations, but due to discrimination, women are sorted into lower ranked

* occupations, conditional on ability. Also assumes that in the absence of

* discrimination, women and men would sort into identical occupations (on average).

generate occupation = (1) + (2)*ability + (0)*female + (-2)*discrimination
+ rnormal()

* The wage is a function of discrimination even in identical jobs, occupational

* choice (which is also affected by discrimination) and ability.

generate wage = (1) + (-1)*discrimination + (1)*occupation + 2*ability + rnormal()

* Assume that ability is unobserved. Then if we regress female on wage, we get a

* a consistent estimate of the unconditional effect of discrimination—
i.e.,

* both the direct effect (paying women less in the same job) and indirect effect

* (occupational choice).

regress wage female

* But occupational choice is correlated with the unobserved factor ability *and*

* it is correlated with female, so renders our estimate on female and occupation

* no longer informative.

regress wage female occupation

* Of course, if we could only control for ability...

regress wage female occupation ability

Examples of collider bias #2: qualitative change in sign Sometimes the problem with conditioning on a collider, though, can be so severe that the correlation becomes statistically insignificant, or worse, even switches sign. Let's see an example where that is true.

Covariates:	Biased unconditional	Biased	Unbiased conditional
Female	-3.074*** (0.000)	0.601*** (0.000)	-0.994*** (0.000)
Occupation		1.793*** (0.000)	0.991*** (0.000)
Ability			2.017*** (0.000)
N	10,000	10,000	10,000
Mean of dependent variable	0.45	0.45	0.45

Table 8: Regressions illustrating confounding bias with simulated gender disparity

```

clear all
set seed 541

* Creating collider bias
* Z -> D -> Y
* D ->X <- Y

* 2500 independent draws from standard normal distribution
clear
set obs 2500
gen z = rnormal()
gen k = rnormal(10,4)
gen d = 0
replace d =1 if k>=12

* Treatment effect = 50. Notice y is not a function of X.
gen y = d*50 + 100 + rnormal()

gen x = d*50 + y + rnormal(50,1)

* Regression
reg y d, robust
reg y x, robust
reg y d x, robust

```

Covariates:	1	2	3
d	50.004*** (0.044)		-0.757 (1.024)
x		0.500*** (0.000)	0.508*** (0.010)
N	2,500	2,500	2,500
Mean of dependent variable	114.90	114.90	114.90

Table 9: Regressions illustrating collider bias

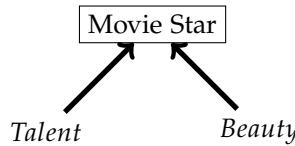
Okay, so let's walk through this exercise. We can see from the above code that the treatment effect is 50, because we coded y as gen

$y = d*50 + 100 + rnormal()$. It is for this reason when we run the first regression, we get a coefficient of 49.998 (column 1). Next we ran a regression of Y on X . Here when we do this, we find a significant effect, yet recall that Y is not a function of X . Rather, X is a function of Y . So this is a spurious result driven by reverse causality. That said, surely we can at least control for X in a regression of Y on D , right? Column 3 shows the impossibility of this regression; it makes it impossible to recover the causal effect of D on Y when we control for X . Why? Because X is a collider, and by conditioning on it, we are introducing new systematic correlations between D and Y that are wiping out the causal effect.

Examples of collider bias: Nonrandom sample selection Maybe this is still not clear. I hope that the following example, therefore, will clarify matters, as it will end in a picture and a picture speaks a thousand words.

A 2009 CNN.com article stated that Megan Fox, of Transformers, was voted the worst and most attractive actress of 2009. While not explicit in the article, the implication of the article was that talent and beauty were negatively correlated. But are they? What if they are in fact independent of each other, but the negative correlation found is a result of a collider bias? What would that look like?⁴⁴

To illustrate, we will generate some data based on the following DAG:



Run the following program in Stata.

```

clear all
set seed 3444

* 2500 independent draws from standard normal distribution
set obs 2500
generate beauty=rnormal()
generate talent=rnormal()

* Creating the collider variable (star)
gen score=(beauty+talent)
egen c85=pctile(score), p(85)
gen star=(score>=c85)
label variable star "Movie star"

* Conditioning on the top 15%
  
```

⁴⁴ I wish I had thought of this example, but alas, I didn't. Gabriel Rossman gets full credit.

```
twoway (scatter beauty talent, mcolor(black) msize(small)
msymbol(smx)),
ytitle(Beauty) xtitle(Talent) subtitle(Aspiring actors and
actresses) by(star, total)
```



Figure 7: Top left figure: Non-star sample scatter plot of beauty (vertical axis) and talent (horizontal axis). Top right figure: Star sample scatter plot of beauty and talent. Bottom left figure: Entire (stars and non-stars combined) sample scatter plot of beauty and talent.

The bottom left panel shows the scatterplot between talent and beauty. Notice that the two variables are independent draws from the standard normal distribution, creating an oblong data cloud. But, because “movie star” is in the top 15 percentile of the distribution of a linear combination of talent and beauty, the movie star sample is formed by a frontier of the combined variables. This frontier has a negative slope and is in the upper right portion of the data cloud, creating a negative correlation between the observations in the movie star sample. Likewise, the collider bias has created a negative correlation between talent and beauty in the non-movie star sample as well. Yet we know that there is in fact *no* relationship between the two variables. This kind of sample selection creates spurious correlations.⁴⁵

Conclusion In conclusion, DAGs are powerful tools. There is far more to them than I have covered here. If you are interested in learning more about them, then I encourage you to carefully read Pearl [2009], which is his magnum opus. It’s a major contribution to the theory of causation, and in my opinion, his ideas merit inclusion in your toolkit as you think carefully about identifying causal effects

⁴⁵ A random sample of the full population would be sufficient to show that there is no relationship between the two variables.

with observational data. DAGs are helpful at both clarifying the relationships between variables, but more importantly than that, DAGs make explicit whether you can identify a causal effect in your dataset. The concept of the backdoor criterion is one way by which you can hope to achieve that identification, and DAGs will help guide you to the identification strategy that satisfies that criterion. Finally, I have found that students learn a lot through this language of DAGs, and since Pearl [2009] shows that DAGs subsume the potential outcomes model (more on that in the next chapter), you need not worry that it is creating unnecessary complexity and contradictions in your pedagogy.

Potential outcomes causal model

Practical questions about causation has been a preoccupation of economists for several centuries. Adam Smith wrote about the causes of the wealth of nations [Smith, 2003]. Karl Marx was interested in the transition of society from capitalism to socialism [Needleman and Needleman, 1969]. The 20th century Cowles Commission sought to better understand the identification problem [Heckman and Vytlacil, 2007].⁴⁶

We can see the development of the modern causality concepts in the writings of several philosophers. Hume [1993] described causation as sequence of temporal events in which had the first event not occurred, the subsequent ones would not either. An example of this is where he said:

“[w]e may define a cause to be an object, followed by another, and where all the objects similar to the first are followed by objects similar to the second. Or in other words where, if the first object had not been, the second never had existed”

Mill [2010] devised five methods for inferring causation. Those methods were (1) the method of agreement, (2) the method of difference, (3) the joint method, (4) the method of concomitant variation and (5) the method of residues. The second method, the method of differences, is most similar to the idea of causation as a comparison among counterfactuals. For instance, he wrote:

“If a person eats of a particular dish, and dies in consequence, that is, would not have died if he had not eaten it, people would be apt to say that eating of that dish was the source of his death.”

Statistical inference A major jump in our understanding of causation occurs coincident with the development of modern statistics. Probability theory and statistics revolutionized science in the 19th century, originally with astronomy. Giuseppe Piazzi, an early 19th century astronomer, discovered the dwarf planet Ceres, located between Jupiter and Mars, in 1801. Piazzi observed it 24 times before it was lost again. Carl Friedrich Gauss proposed a method which

⁴⁶ This brief history will focus on the development of the potential outcomes model. See Morgan [1991] for a more comprehensive history of econometric ideas.

could successfully predict Ceres' next location using data on its prior location. His method minimized the sum of the squared errors, or *ordinary least squares*. He discovered it at age 18 and published it in 1809 at age 24 [Gauss, 1809]. Other contributors include LaPlace and Legendre.

Regression analysis enters the social sciences through the work of statistician G. Udny Yule. Yule [1899] was interested in the causes of poverty in England. Poor people depended on either poor-houses or the local authorities. Yule wanted to know if public assistance increased the number of paupers, which is a causal question. Yule used Gauss's least squares method to estimate the partial correlation between public assistance and poverty. Here was his data, drawn from the English Censuses of 1871 and 1881. Download it using scuse.

```
. scuse yule
```

Each row is a particular location in England (e.g., Chelsea, Strand). And the second through fourth columns are growth rates. Yule estimated a model similar to the following:

$$Pauper = \beta_0 + \beta_1 Outrelief + \beta_2 Old + \beta_3 Pop + u$$

Using our data, we would estimate this using the regress command:

```
. regress paup outrelief old pop
```

His results are reported in Table 13.

Covariates	Dependent variable
	Pauperism growth
Outrelief	0.752 (0.135)
Old	0.056 (0.223)
Pop	-0.311 (0.067)

Table 10: Yule regressions [Yule, 1899].

In words, a 10 percentage point change in the outrelief growth rate is associated with a 7.5 percentage point increase in the pauperism growth rate, an elasticity of 0.75. Yule used regression to isolate the effects of out-relief, and his principal conclusion was that welfare increased pauper growth rates. What's wrong with his statistical reasoning? Do we think that the unobserved determinants of pauperism

growth rates are uncorrelated with out-relief growth rates? After all, he does not control for any economic factors which surely affect both poverty and the amount of resources allocated to out-relief. Likewise, he may have the causality backwards – perhaps the growth in pauperism is the cause of the growth in out-relief, not the other way around. But, despite its flaws, it represented the first known instance where statistics (and regression in particular) was used to estimate a policy-relevant causal effect.

Physical randomization The notion that physical randomization was the foundation of causal inference was in the air in the 19th and early 20th century, but it was not until Fisher [1935] that it crystalized. The first historically recognized randomized experiment was fifty years earlier in psychology [Peirce and Jastrow, 1885]. But interestingly, their reason for randomization was *not* as the basis for causal inference. Rather, they proposed randomization as a way of fooling subjects in their experiments. Peirce and Jastrow [1885] were using an experiment on subjects that had a sequence of treatments, and they used physical randomization so that participants couldn't guess at what would happen next. But Peirce appears to have anticipated Neyman's concept of unbiased estimation when using random samples and appears to have even thought of randomization as a physical process to be implemented in practice, but no one can find any suggestion for the physical randomization of treatments to units as a basis for causal inference until Splawa-Neyman [1923] and Fisher [1925].

Splawa-Neyman [1923] develops the very useful potential outcomes notation, and while he proposes randomization, it is not taken to be literally necessary until Fisher [1925]. Fisher [1925] proposes the explicit use of randomization in experimental design for causal inference.⁴⁷

Fisher [1935] described a thought experiment in which a lady claims she can discern whether milk or tea was poured first in a cup of tea. While he does not give her name, we now know that the lady in the thought experiment was Muriel Bristol and that the thought experiment in fact did happen.⁴⁸ Muriel Bristol established the Rothamstead Experiment Station in 1919 and was a PhD scientist back in the days when women weren't PhD scientists. One day during afternoon tea, Muriel claimed that she could tell whether the milk was added to the cup before or after the tea, which as one might guess, got a good laugh from her male colleagues. Fisher took the bait and devised the following randomized experiment.

Given a cup of tea with milk, a lady claims she can discern whether milk or tea was first added to the cup. To test her claim,

⁴⁷ For more on the transition from Splawa-Neyman [1923] to Fisher [1925], see Rubin [2005].

⁴⁸ Apparently, Bristol correctly guessed all four cups of tea.

8 cups of tea were prepared, 4 of which the milk was added first, and 4 where the tea was added first. How many cups does she have to correctly identify to convince us of her uncanny ability?

Fisher [1935] proposed a kind of permutation-based inference – a method we now call the Fisher exact test. She possesses the ability probabilistically, not with certainty, if the likelihood of her guessing all four correctly was sufficiently low. There are $8 \times 7 \times 6 \times 5 = 1,680$ ways to choose a first cup, a second cup, a third cup, and a fourth cup, in order. There are $4 \times 3 \times 2 \times 1 = 24$ ways to order 4 cups. So the number of ways to choose 4 cups out of 8 is $\frac{1680}{24} = 70$. Note, the lady performs the experiment by selecting 4 cups. The probability that she would correctly identify all 4 cups is $\frac{1}{70}$. Either she has no ability, and has chosen the correct 4 cups by chance alone, or she has the discriminatory ability that she claims. Since choosing correctly is highly unlikely (one chance in 70), we decide for the second.

To only get 3 right, she would have to choose 3 from the 4 correct ones. She can do this by $4 \times 3 \times 2 = 24$ with order. Since 3 cups can be ordered in $3 \times 2 = 6$ ways, there are 4 ways for her to choose 3 correct. Since she can now choose 1 incorrect cup 4 ways, there are a total of $4 \times 4 = 16$ ways for her to choose exactly 3 right and 1 wrong. Hence the probability that she chooses exactly 3 correctly is $\frac{16}{70}$. If she got only 3 correct and 1 wrong, this would be evidence for her ability, but not persuasive evidence, since getting 3 correct is $\frac{16}{70} = 0.23$.

Causal inference, in this context, is a probabilistic idea wherein the observed phenomena is compared against permutation-based randomization called the *null hypothesis*. The null hypothesis is a specific description of a possible state of nature. In this example, the null hypothesis is that the lady has no special ability to discern the order in which milk is poured into tea, and thus, the observed phenomena was only by chance. We can never prove the null, but the data may provide evidence to reject it. In most situations, we are trying to reject the null hypothesis.

Medicine and Economics Physical randomization had largely been the domain of agricultural experiments until the mid-1950s when it began to be used in medical trials. One of the first major randomized experiments in medicine were polio vaccination trials. The Salk polio vaccine field trials was one of the largest randomized experiments ever attempted, as well as one of the earliest. In 1954, the Public Health Service set out to answer whether the Salk vaccine prevented polio. Children in the study were assigned *at random* to receive the vaccine or a placebo.⁴⁹ Also the doctors making the diagnoses of polio did not know whether the child had received the vaccine or the placebo. The polio vaccine trial was a *double-blind, randomized*

⁴⁹ In the placebo, children were inoculated with a saline solution.

controlled trial. It was necessary for the field trial to be very large because the rate at which polio occurred in the population was 50 per 100,000. The treatment group, which contained 200,745 individuals, saw 33 polio cases. The control group who had been inoculated had 201,229 individuals, and saw 115 cases. The probability of seeing this big a difference by chance alone is about 1 in a billion. The only plausible explanation, it was argued, was that the polio vaccine caused a reduction in the risk of polio.

A similar large scale randomized experiment occurred in economics in the 1970s. Between 1971 and 1982, the Rand corporation conducted a large-scale randomized experiment studying the causal effect of healthcare insurance on healthcare utilization. For the study, Rand recruited 7,700 individuals under age 65. The experiment was somewhat complicated with multiple treatment arms. Participants were randomly assigned to one of five health insurance plans: free care, three types with varying levels of cost sharing, and an HMO plan. Participants with cost sharing made fewer physician visits and had fewer hospitalizations than those with free care. Other declines in health care utilization, such as we fewer dental visits, were also found among the cost-sharing treatment groups. Overall, participants in the cost sharing plans tended to spend less on health which came from using fewer services. The reduced use of services occurred mainly because participants in the cost sharing treatment groups were opting not to initiate care.⁵⁰

Potential outcomes While the potential outcomes ideas were around, it did not become the basis of causal inference in the social sciences until Rubin [1974].⁵¹ In the potential outcomes tradition [Splawa-Neyman, 1923, Rubin, 1974], a causal effect is defined as a comparison between two states of the world. In the first state of the world, a man takes aspirin for his headache and one hour later reports the severity of his headache. In the second state of the world, that same man refused aspirin and one hour later reported the severity of his headache. What was the causal effect of the aspirin? According to Rubin, the causal effect of the aspirin is the difference in the severity of his headache between two states of the world: one where he took the aspirin (the actual state of the world) and one where he never took the aspirin (the counterfactual state of the world). The difference between these two dimensions, if you would, at the same point in time represents the causal effect of the intervention itself.

To ask questions like this is to engage in a kind of storytelling. Humans have always been interested in stories exploring counterfactuals. Examples include Christmas Carol, It's a Wonderful Life and Man in the High Castle, just to name just a few. What if Bruce

⁵⁰ More information about this fascinating experiment can be found in Newhouse [1993].

⁵¹ The idea of causation as based on counterfactuals appears in philosophy independent of Rubin [1974] with Lewis [1973]. Some evidence for it may exist in John Stuart Mill's methods for causal inference as well.

Wayne's parents had never been murdered? What if that waitress had won the lottery? What if your friend from high school had never taken that first drink? What if Neo had taken the blue pill? These are the sort of questions that can keep a person up at night.

But it's important to note that these kinds of questions are by definition *unanswerable*.⁵² To wonder how life would be different had one single event been changed is to indulge in counterfactual reasoning, and since counterfactuals by definition don't exist, the question cannot be answered. History is a sequence of observable, *factual* events, one after another. We don't know what would have happened had one event changed because we are missing data on the *counterfactual*. Potential outcomes exist *ex ante* as a set of possibilities, but once a decision is made, all but one of them disappears.

Donald Rubin, and statisticians Roland Fisher and Jerzy Neyman before him, take as a starting point that a causal effect is a comparison across two potential outcomes.⁵³ To make this concrete, we introduce some notation and language. For simplicity, we will assume a *dummy* variable that takes on a value of one if a particular unit i receives the *treatment* and a zero if they do not.⁵⁴ Each unit will have two *potential outcomes*, but only one observed outcome. Potential outcomes are defined as Y_i^1 if the unit received the treatment and Y_i^0 if the unit did not. We'll call the state of the world where no treatment occurred the *control* state. Notice the superscripts and the subscripts – each unit i has exactly two potential outcomes: a potential outcome under a state of the world where the treatment occurred (Y^1) and a potential outcome where the treatment did not occur (Y^0).

Observable outcomes, Y_i , are distinct from potential outcomes. Whereas potential outcomes are hypothetical random variables that differ across the population, observable outcomes are factual random variables. A unit's observable outcome is determined according to a *switching equation*:

$$Y_i = D_i Y_i^1 + (1 - D_i) Y_i^0 \quad (76)$$

where D_i equals one if the unit received the treatment and zero if it did not. Notice the logic of the equation. When $D_i = 1$, then $Y_i = Y_i^1$ because the second term zeroes out. And when $D_i = 0$, the first term zeroes out and therefore $Y_i = Y_i^0$.

Rubin defines a treatment effect, or causal effect, as simply the difference between two states of the world:

$$\delta_i = Y_i^1 - Y_i^0$$

Immediately we are confronted with a problem. If a treatment effect requires knowing two states of the world, Y_i^1 and Y_i^0 , but by the switching equation we only observe one, then we cannot calculate the treatment effect.

⁵² It is also worth noting that counterfactual reasoning appears to be a hallmark of the human mind. We are unusual among creatures in that we are capable of asking and imagining these types of what-if questions.

⁵³ This analysis can be extended to more than two potential outcomes, but for simplicity we will stick with just two.

⁵⁴ The treatment here is any particular intervention, or causal variable of interest. In economics, it is usually the comparative statics exercise.

Average treatment effects From this simple definition of a treatment effect come three different parameters that are often of interest to researchers. They are all population means. The first is called the *average treatment effect* and it is equal to

$$E[\delta_i] = E[Y_i^1 - Y_i^0] = E[Y_i^1] - E[Y_i^0]$$

Notice, as with our definition of individual level treatment effects, the average treatment effect is unknowable as well because it requires two observations per unit i , one of which is a counterfactual. Thus the average treatment effect, *ATE*, like the individual treatment effect, is not a quantity that can be calculated with any data set known to man.

The second parameter of interest is the *average treatment effect for the treatment group*. That's a mouthful, but let me explain. There exist two groups of people: there's a treatment group and there's a control group. The average treatment effect for the treatment group, or *ATT* for short, is simply that population mean treatment effect for the group of units that have been assigned the treatment in the first place. Insofar as δ_i differs across the population, the ATT may be different from the ATE. In observational data, it almost always will be in fact different from the ATE. And, like the ATE, it is unknowable, because like the ATE, it requires two observations per treatment unit i :

$$\begin{aligned} ATT &= E[\delta_i | D_i = 1] \\ &= E[Y_i^1 - E_i^0 | D_i = 1] \\ &= E[Y_i^1 | D_i = 1] - E[Y_i^0 | D_i = 1] \end{aligned}$$

The final parameter of interest is called the average treatment effect for the control group, or *untreated* group. Its shorthand is *ATU* which stands for average treatment effect for the untreated. And like its ATT brother, the ATU is simply the population mean treatment effect for the units in the control group. Given heterogeneous treatment effects, it's probably the case that the $ATT \neq ATU$ – especially in an observational setting. The formula for the ATU is

$$\begin{aligned} ATU &= E[\delta_i | D_i = 0] \\ &= E[Y_i^1 - Y_i^0 | D_i = 0] \\ &= E[Y_i^1 | D_i = 0] - E[Y_i^0 | D_i = 0] \end{aligned}$$

Depending on the research question, one or all three of these parameters are interesting. But the two most common ones of interest are the *ATE* and the *ATT*.

Simple difference in means decomposition This has been somewhat abstract, so let's be concrete. Let's assume there are ten patients i

who have cancer, and two medical procedures or treatments. There is a surgery intervention, $D_i = 1$, and there is a chemotherapy intervention, $D_i = 0$. Each patient has the following two potential outcomes where a potential outcome is defined as post-treatment lifespan in years:

Patients	Y^1	Y^0	δ
1	7	1	6
2	5	6	-1
3	5	1	4
4	7	8	-1
5	4	2	2
6	10	1	9
7	1	10	-9
8	5	6	-1
9	3	7	-4
10	9	8	1

Table 11: Potential outcomes for ten patients receiving surgery Y^1 or chemo Y^0 .

We can calculate the average treatment effect if we have this matrix of data because the average treatment effect is simply the mean difference between columns 2 and 3. That is $E[Y^1] = 5.6$ and $E[Y^0] = 5$, which means that $ATE = 0.6$. In words, the average causal effect of surgery for these ten patients is 0.6 additional years (compared to chemo).⁵⁵

Now notice carefully: not everyone benefits from surgery. Patient 7, for instance, lives only 1 additional year post-surgery versus 10 additional years post-chemo. But the ATE is simply the average over these heterogeneous treatment effects.

To maintain this fiction, let's assume that there exists the perfect doctor.⁵⁶ The perfect doctor knows each person's potential outcomes and chooses the treatment that is best for each person. In other words, he chooses to put them in surgery or chemotherapy depending on whichever treatment has the longer post-treatment lifespan. Once he makes that treatment assignment, he observes their post-treatment actual outcome according to the switching equation we mentioned earlier.

Table 12 differs from Table 11 because Table 11 shows only the potential outcomes, but Table 12 shows only the *observed* outcome for treatment and control group. Once treatment has been assigned, we can calculate the average treatment effect for the surgery group (ATT) versus the chemo group (ATU). The ATT equals 4.4 and the ATU equals -3.2. In words, that means that the average post-surgery lifespan for the surgery group is 4.4 additional years, whereas the

⁵⁵ Note that causality always involves comparisons.

⁵⁶ I credit Donald Rubin with this example [Rubin, 2004]

Patients	Y	D
1	7	1
2	6	0
3	5	1
4	8	0
5	4	1
6	10	1
7	10	0
8	6	0
9	7	0
10	9	1

Table 12: Post-treatment observed lifespans in years for surgery $D = 1$ versus chemotherapy $D = 0$.

average post-surgery lifespan for the chemotherapy group is 3.2 fewer years.⁵⁷

Now the ATE is 0.6, which is just a weighted average between the ATT and the ATU.⁵⁸ So we know that the overall effect of surgery is positive, though the effect for some is negative. There exist heterogeneous treatment effects in other words, but the net effect is positive. But, what if we were to simply compare the average post-surgery lifespan for the two groups? This is called an *estimate* of the ATE – it takes observed values, calculates means, in an effort to *estimate* the parameter of interest, the ATE. We will call this simple difference in mean outcomes the SDO,⁵⁹ and it is simply equal to

$$E[Y^1|D = 1] - E[Y^0|D = 0]$$

which can be estimated using samples of data

$$\begin{aligned} SDO &= E[Y^1|D = 1] - E[Y^0|D = 0] \\ &= \frac{1}{N_T} \sum_{i=1}^n (y_i|d_i = 1) - \frac{1}{N_C} \sum_{i=1}^n (y_i|d_i = 0) \end{aligned}$$

which in this situation is equal to $7 - 7.4 = -0.4$. Or in words, the treatment group lives 0.4 fewer years post-surgery than the chemo group. Notice how misleading this statistic is, though. We know that the average treatment effect is positive, but the simple difference in mean outcomes is negative. Why is it different? To understand why it is different, we will decompose the simple difference in mean outcomes using LIE and the definition of ATE. Note there are three parts to the SDO. Think of the left hand side as the calculated average, but the right hand side as the truth about that calculated average.

$$\begin{aligned} E[Y^1|D = 1] - E[Y^0|D = 0] &= ATE \\ &\quad + E[Y^0|D = 1] - E[Y^0|D = 0] \\ &\quad + (1 - \pi)(ATT - ATU) \end{aligned} \tag{77}$$

⁵⁷ The reason that the ATU is negative is because the treatment here is the surgery, which was the worse treatment of the two of them. But you could just as easily interpret this as 3.2 *additional* years of life if they had received chemo instead of surgery.

⁵⁸ $ATE = p \times ATT + (1 - p) \times ATU = 0.5 \times 4.4 + 0.5 \times -3.2 = 0.6$.

⁵⁹ Morgan and Winship [2014] call this estimator the *naive average treatment effect* or NATE for short.

To understand where these parts on the right-hand-side originate, we need to start over and decompose the parameter of interest, ATE , into the sum of four parts using the law of iterated expectations. ATE is equal to sum of conditional average expectations, ATT and ATU , by LIE

$$\begin{aligned} ATE &= E[Y^1] - E[Y^0] \\ &= \{\pi E[Y^1|D = 1] + (1 - \pi)E[Y^1|D = 0]\} \\ &\quad - \{\pi E[Y^0|D = 1] + (1 - \pi)E[Y^0|D = 0]\} \end{aligned}$$

where π is the share of patients who received surgery and $1 - \pi$ is the share of patients that received chemotherapy. Because the conditional expectation notation is a little cumbersome, let's exchange each term on the left hand side, ATE , and right hand side, the part we got from LIE, using some letters. This will allow the proof to be a little less cumbersome to follow.

$$\begin{aligned} E[Y^1|D = 1] &= a \\ E[Y^1|D = 0] &= b \\ E[Y^0|D = 1] &= c \\ E[Y^0|D = 0] &= d \\ ATE &= e \end{aligned}$$

Now through the following algebraic manipulation.

$$\begin{aligned} e &= \{\pi a + (1 - \pi)b\} - \{\pi c + (1 - \pi)d\} \\ e &= \pi a + b - \pi b - \pi c - d + \pi d \\ e &= \pi a + b - \pi b - \pi c - d + \pi d + (\mathbf{a} - \mathbf{a}) + (\mathbf{c} - \mathbf{c}) + (\mathbf{d} - \mathbf{d}) \\ 0 &= e - \pi a - b + \pi b + \pi c + d - \pi d - \mathbf{a} + \mathbf{a} - \mathbf{c} + \mathbf{c} - \mathbf{d} + \mathbf{d} \\ \mathbf{a} - \mathbf{d} &= e - \pi a - b + \pi b + \pi c + d - \pi d + \mathbf{a} - \mathbf{c} + \mathbf{c} - \mathbf{d} \\ \mathbf{a} - \mathbf{d} &= e + (\mathbf{c} - \mathbf{d}) + \mathbf{a} - \pi a - b + \pi b - \mathbf{c} + \pi c + d - \pi d \\ \mathbf{a} - \mathbf{d} &= e + (\mathbf{c} - \mathbf{d}) + (1 - \pi)a - (1 - \pi)b + (1 - \pi)d - (1 - \pi)c \\ \mathbf{a} - \mathbf{d} &= e + (\mathbf{c} - \mathbf{d}) + (1 - \pi)(a - c) - (1 - \pi)(b - d) \end{aligned}$$

Now substituting our definitions, we get the following:

$$\begin{aligned} E[Y^1|D = 1] - E[Y^0|D = 0] &= ATE \\ &\quad + (E[Y^0|D = 1] - E[Y^0|D = 0]) \\ &\quad + (1 - \pi)(ATT - ATU) \end{aligned}$$

And the proof ends. Now the left hand side can be estimated with a

sample of data. And the right-hand-side is equal to the following:

$$\underbrace{\frac{1}{NT} \sum_{i=1}^n (y_i | d_i = 1) - \frac{1}{NC} \sum_{i=1}^n (y_i | d_i = 0)}_{SDO} = \underbrace{E[Y^1] - E[Y^0]}_{\text{Average Treatment Effect}} + \underbrace{E[Y^0 | D = 1] - E[Y^0 | D = 0]}_{\text{Selection bias}} + \underbrace{(1 - \pi)(ATT - ATU)}_{\text{Heterogenous treatment effect bias}}$$

Let's discuss each of these in turn. The left-hand-side is the simple difference in mean outcomes and we know it is equal to -0.4 . Thus it must be the case that the right hand side sums to -0.4 . The first term is the average treatment effect, which is the parameter of interest. We know that it is equal to $+0.6$. Thus the remaining two terms must be the source of the bias that is causing our $SDO < ATE$. The second term is called the *selection bias* which merits some unpacking. The selection bias is the inherent differences between the two groups if they both received chemo. Usually, though, it's just a description of the differences between the two if there had never been a treatment in the first place. There are in other words two groups: there's a surgery group and there's a chemo group. How do their potential outcomes under control differ? Notice that the first is a counterfactual, whereas the second is an observed outcome according to the switching equation. We can calculate this difference here because we have the complete potential outcomes in Table 11. That difference is equal to -4.8 . The third term is a lesser known form of bias, but we include it to be comprehensive, and because we are focused on the ATE.⁶⁰ The *heterogenous treatment effect bias* is simply the different returns to surgery for the two groups multiplied by the share of the population that is in the chemotherapy group at all. This final term is $0.5 \times (4.4 - (-3.2))$ which is 3.8 . Note in case it's not obvious, the reason that $\pi = 0.5$ is because 5 units out of 10 units are in the chemotherapy group.

Now that we have all three parameters on the right-hand-side, we can see why the SDO is equal to -0.4 .

$$-0.4 = 0.6 - 4.8 + 3.8$$

Notice that the SDO actually does contain the parameter of interest. But the problem is that that parameter of interest is confounded by two forms of bias, the selection bias and the heterogeneous treatment effect bias. If there is a constant treatment effect, $\delta_i = \delta \forall i$, then $ATU = ATT$ and so $SDO = ATE + \text{selection bias}$. A large part of empirical research is simply trying to develop a strategy for eliminating selection bias.

⁶⁰ Note that Angrist and Pischke [2009] have a slightly different decomposition where the $SDO = ATT + \text{selection bias}$, but that is because their parameter of interest is the ATT and therefore the third term doesn't appear.

Let's start with the most credible situation for using *SDO* to estimate *ATE*: when the treatment itself (e.g., surgery) has been assigned to patients *independent* of their potential outcomes. Notationally speaking, this is

$$(Y^1, Y^0) \perp\!\!\!\perp D$$

Now in our example, we already know that this is violated because the perfect doctor specifically chose surgery or chemo based on their potential outcomes. Specifically, they received surgery if $Y^1 > Y^0$ and chemo if $Y^1 < Y^0$. Thus in our case, the perfect doctor ensured that D depended on Y^1, Y^0 .

But, what if he hadn't done that? What if he had chosen surgery in such a way that did not depend on Y^1 or Y^0 ? What might that look like? For instance, maybe he alphabetized them by last name, and the first five received surgery and the last five received chemotherapy. Or maybe he used the second hand on his watch to assign surgery to them: if it was between 1 – 30 seconds, he gave them surgery, and if it was between 31 – 60 seconds he gave them chemotherapy.⁶¹ In other words, let's say that he chose some method for assigning treatment that did not depend on the values of potential outcomes under either state of the world. What would that mean in this context? Well, it would mean that

$$\begin{aligned} E[Y^1|D = 1] - E[Y^1|D = 0] &= 0 \\ E[Y^0|D = 1] - E[Y^0|D = 0] &= 0 \end{aligned}$$

Or in words, it would mean that the mean potential outcome for Y^1 or Y^0 is the same (in the population) for either the surgery group or the chemotherapy group. This kind of *randomization* of the treatment assignment would eliminate both the selection bias and the heterogeneous treatment effect bias. Let's take it in order. The selection bias zeroes out as follows:

$$E[Y^0|D = 1] - E[Y^0|D = 0] = 0$$

And thus the *SDO* no longer suffers from selection bias. How does randomization affect heterogeneity treatment bias from the third line? Rewrite definitions for ATT and ATU:

$$\begin{aligned} \text{ATT} &= E[Y^1|D = 1] - E[Y^0|D = 1] \\ \text{ATU} &= E[Y^1|D = 0] - E[Y^0|D = 0] \end{aligned}$$

⁶¹ In Craig [2006], a poker-playing banker used his watch as a random number generator to randomly bluff in certain situations.

Rewrite the third row bias after $1 - \pi$:

$$\begin{aligned} ATT - ATU &= E[Y^1 | D=1] - E[Y^0 | D=1] \\ &\quad - E[Y^1 | D=0] + E[Y^0 | D=0] \\ &= 0 \end{aligned}$$

If treatment is independent of potential outcomes, then:

$$\begin{aligned} \frac{1}{N_T} \sum_{i=1}^n (y_i | d_i = 1) - \frac{1}{N_C} \sum_{i=1}^n (y_i | d_i = 0) &= E[Y^1] - E[Y^0] \\ SDO &= ATE \end{aligned}$$

What's necessary in this situation is simply (a) data on observable outcomes, (b) data on treatment assignment, and (c) $(Y^1, Y^0) \perp\!\!\!\perp D$. We call (c) the independence assumption. To illustrate that this would lead to the SDO, we will use the following Monte Carlo simulation. Note that ATE in this example is equal to 0.6.

```
clear all
program define gap, rclass
version 14.2
syntax [, obs(integer 1) mu(real 0) sigma(real 1) ]
clear
drop _all
set obs 10
gen y1 = 7 in 1
replace y1 = 5 in 2
replace y1 = 5 in 3
replace y1 = 7 in 4
replace y1 = 4 in 5
replace y1 = 10 in 6
replace y1 = 1 in 7
replace y1 = 5 in 8
replace y1 = 3 in 9
replace y1 = 9 in 10
gen y0 = 1 in 1
replace y0 = 6 in 2
replace y0 = 1 in 3
replace y0 = 8 in 4
replace y0 = 2 in 5
replace y0 = 1 in 6
replace y0 = 10 in 7
replace y0 = 6 in 8
replace y0 = 7 in 9
replace y0 = 8 in 10
```

```

drawnorm random
sort random
gen d=1 in 1/5
replace d=0 in 6/10
gen y=d*y1 + (1-d)*y0
egen sy1 = mean(y) if d==1
egen sy0 = mean(y) if d==0
collapse (mean) sy1 sy0
gen sdo = sy1 - sy0
keep sdo
summarize sdo
gen mean = r(mean)
end
simulate mean, reps(10000): gap
su _sim_1

```

This Monte Carlo runs 10,000 times, each time calculating the average SDO under independence – which is ensured by the random number sorting that occurs. In my running of this program, the ATE is 0.6 and the SDO is on average equal to 0.59088.⁶²

Before we move on from the SDO, let's just re-emphasize something that is often lost on students first learning the independence concept and notation. Independence does not imply that $E[Y^1|D = 1] - E[Y^0|D = 0] = 0$. Nor does it imply that $E[Y^1|D = 1] - E[Y^0|D = 1] = 0$. Rather, it implies

$$E[Y^1|D = 1] - E[Y^1|D = 0] = 0$$

in a large population. That is, independence implies that the two groups of units, surgery and chemo, have the same potential outcome on average in the population.

How realistic is independence in observational data? Economics – maybe more than any other science – tells us that independence is unlikely to hold observationally. Economic actors are always attempting to achieve some optima. For instance, parents are putting kids in what they perceive to be the best school for them and that is based on potential outcomes. In other words, people are *choosing* their interventions and most likely that decision is related to the potential outcomes, which makes simple comparisons improper. Rational choice is always pushing against the independence assumption, and therefore simple comparison in means will not approximate the true causal effect. We need unit randomization for simple comparisons to tell us anything meaningful.

One last thing. Rubin argues that there are a bundle of assumptions behind this kind of calculation, and he calls these assumptions

⁶² Because it's not seeded, when you run it, your answer will be close but slightly different due to the randomness of the sample drawn.

the *stable unit treatment value assumption* or SUTVA for short. That's a mouthful, but here's what it means. It means that we are assuming the unit-level treatment effect ("treatment value") is fixed over the entire population, which means that the assignment of the treatment to one unit cannot affect the treatment effect or the potential outcomes of another unit.

First, this implies that the treatment is received in homogenous doses to all units. It's easy to imagine violations of this though – for instance if some doctors are better surgeons than others. In which case, we just need to be careful what we are and are not defining as the treatment.

Second, this implies that there are no externalities, because by definition, an externality spills over to other units untreated. In other words, if unit 1 receives the treatment, and there is some externality, then unit 2 will have a different Y^0 value than she would have if unit 1 had not received the treatment. We are assuming away this kind of spillover.

Related to that is the issue of general equilibrium. Let's say we are estimating the causal effect of returns to schooling. The increase in college education would in general equilibrium cause a change in relative wages that is different than what happens under partial equilibrium. This kind of scaling up issue is of common concern when one consider extrapolating from the experimental design to the large scale implementation of an intervention in some population.

STAR Experiment Now I'd like to discuss a large-scale randomized experiment to help explain some of these abstract concepts. Krueger [1999] analyzed a 1980s randomized experiment in Tennessee called the Student/Teacher Achievement Ratio (STAR). This was a state-wide randomized experiment that measured the average treatment effect of class size on student achievement. There were two arms to the treatment: a small class of 13-17 students, and a regular sized classroom of 22-25 students with a full-time teacher's aide. The control group was a regular sized classroom of 22-25 students with no aide. Approximately 11,600 students and their teachers were *randomly* assigned to one of the three groups. After the assignment, the design called for the students to remain in the same class type for four years (K-3). Randomization occurred within schools at the kindergarten level.

For this section, we will use Krueger's data and attempt to replicate as closely as possible his results. Type in (ignoring the period):

- . clear
- . scuse star_sw

Note that insofar as it was truly a randomized experiment, then the average potential outcomes for students in a small class will be the same as the average potential outcomes for each of the other treatment arms. As such we can simply calculate the mean outcomes for each group and compare them to determine the average treatment effect of a small class size. Nonetheless, it is useful to analyze experimental data with regression analysis because in this instance the randomization was *conditional* on the school itself.

Assume for the sake of argument that the treatment effects are constant. This implies two things: it implies that $Y_i^1 - Y_i^0 = \delta \forall i$, first of all. And second, it implies that $ATE = ATT = ATU$. Thus the simple difference in outcomes SDO is equal to ATE plus selection bias because the heterogenous treatment effect bias zeroes out.

Let's write out the regression equation by first writing out the switching equation:

$$Y_i = D_i Y_i^1 + (1 - D_i) Y_i^0$$

Distributing the Y_i^0 we get

$$Y_i = Y_i^0 + D_i(Y_i^1 - Y_i^0)$$

which is equal to

$$Y_i = Y_i^0 + \delta D_i$$

given the definition of the treatment effect from earlier. Now add $E[Y_i^0] - E[Y_i^0] = 0$ to the right-hand side and rearrange the terms to get

$$Y_i = E[Y_i^0] + \delta D_i + Y_i^0 - E[Y_i^0]$$

and then rewrite as the following regression equation

$$Y_i = \alpha + \delta D_i + u_i$$

where u_i is the random part of Y_i^0 . This is a regression equation we could use to estimate the average treatment effect of D on Y .

We will be evaluating experimental data, and so we could just compare the treatment group to the control group. But we may want to add additional controls in a multivariate regression model for a couple of reasons. The multivariate regression model would be

$$Y_i = \alpha + \delta D_i + X_i \gamma + u_i$$

where X is a matrix of unit specific predetermined covariates unaffected by D . There are two main reasons for including additional controls in the experimental regression model.

1. Conditional random assignment. Sometimes randomization is done *conditional* on some observable. In this example, that's the

school, as they randomized within a school. We will discuss the “conditional independence assumption” later when we cover matching.

2. Additional controls increase precision. Although control variables X_i are uncorrelated with D_i , they may have substantial explanatory power for Y_i . Therefore including them reduces variance in the residuals which lowers the standard errors of the regression estimates.

Krueger estimates the following econometric model

$$Y_{ics} = \beta_0 + \beta_1 SMALL_{cs} + \beta_2 REG/A_{cs} + \alpha_s + \varepsilon_{ics} \quad (78)$$

where i indexes a student, c a class, s a school, Y a student’s percentile score, $SMALL$ a dummy equalling 1 if the student was in a small class, REG/A a dummy equalling 1 if the student was assigned a regular class with an aide and α is a *school fixed effect*. A school fixed effect is simply a dummy variable for each school, and controlling for that means that the variance in $SMALL$ and REG/A is within each school. He did this because the STAR program was randomized within a school – treatment was *conditionally* independent of the potential outcomes.

First, I will produce Krueger’s actual estimates, then I will produce similar regression output using the `star_sw.dta` file we downloaded. Figure 8 shows estimates from this equation using least squares. The first column is a simple regression of the percentile score onto the two treatment dummies with no controls. The effect of a small class is 4.82 which means that the kids in a small class moved 4.82 points up the distribution at the end of year. You can see by dividing the coefficient by the standard error that this is significant at the 5% level. Notice there is no effect of a regular sized classroom with an aide on performance, though. The coefficient is both close to zero and has large standard errors. The R^2 is very small as well – only 1% of the variation in percentile score is explained by this treatment variable.

Columns 2-4 add in additional controls. First Krueger controls for school fixed effects which soaks up a lot of the variation in Y evidenced by the larger R^2 . It’s interesting that the R^2 has increased, but the coefficient estimates have not really changed. This is evidence of randomization. The coefficient on small class size is considerably more precise, but not materially different from what was shown in column 1. This coefficient is also relatively stable when additional student demographics (column 3) and teacher characteristics (column 4) are controlled for. This is a reassuring table in many respects because it is showing that $E[\delta|X] = E[\delta]$ which is an extension of

the independence assumption. That suggests that D is assigned independent of the potential outcomes, because δ is a function of Y^1 and Y^0 .

Explanatory variable	OLS: actual class size			
	(1)	(2)	(3)	(4)
A. Kindergarten				
Small class	4.82 (2.19)	5.37 (1.26)	5.36 (1.21)	5.37 (1.19)
Regular/aide class	.12 (2.23)	.29 (1.13)	.53 (1.09)	.31 (1.07)
White/Asian (1 = yes)	—	—	8.35 (1.35)	8.44 (1.36)
Girl (1 = yes)	—	—	4.48 (.63)	4.39 (.63)
Free lunch (1 = yes)	—	—	-13.15 (.77)	-13.07 (.77)
White teacher	—	—	—	-.57 (2.10)
Teacher experience	—	—	—	.26 (.10)
Master's degree	—	—	—	-.51 (1.06)
School fixed effects	No	Yes	Yes	Yes
R^2	.01	.25	.31	.31

Figure 8: Regression of kindergarten percentile scores onto treatments [Krueger, 1999].

Next Krueger [1999] estimated the effect of the treatments on the first grade outcomes. Note that this is the same group of students from the kindergarten sample, just aged one year. Figure 9 shows the results from this regression. Here we find coefficients that are about 40% larger than the ones we found for kindergarteners. Also, the regular sized classroom with an aide, while smaller, is no longer equal to zero or imprecise. Again, this coefficient is statistically stable across all specifications.

A common problem in randomized experiments involving human beings, that does not plague randomized experiments involving non-humans, is *attrition*. That is, what if people leave the experiment? If attrition is random, then attrition affects the treatment and control groups in the same way (on average). Random attrition means that

Explanatory variable	OLS: actual class size			
	(1)	(2)	(3)	(4)
B. First grade				
Small class	8.57 (1.97)	8.43 (1.21)	7.91 (1.17)	7.40 (1.18)
Regular/aide class	3.44 (2.05)	2.22 (1.00)	2.23 (0.98)	1.78 (0.98)
White/Asian (1 = yes)	—	—	6.97 (1.18)	6.97 (1.19)
Girl (1 = yes)	—	—	3.80 (.56)	3.85 (.56)
Free lunch (1 = yes)	—	—	-13.49 (.87)	-13.61 (.87)
White teacher	—	—	—	-4.28 (1.96)
Male teacher	—	—	—	11.82 (3.33)
Teacher experience	—	—	—	.05 (0.06)
Master's degree	—	—	—	.48 (1.07)
School fixed effects	No	Yes	Yes	Yes
R^2	.02	.24	.30	.30

Figure 9: Regression of first grade percentile scores onto treatments [Krueger, 1999].

our estimates of the average treatment effect remain unbiased.

But in this application, involving schooling, attrition may be non-random. For instance, especially good students placed in large classes may leave the public school for a private school under the belief that large class sizes will harm their child's performance. Thus the remaining students will be people for whom Y^0 is lower, thus giving the impression the intervention was more effective than maybe it actually was. Krueger [1999] addresses this concern by imputing the test scores from their earlier test scores for all children who leave the sample and then re-estimates the model including students with imputed test scores.

TABLE VI
EXPLORATION OF EFFECT OF ATTRITION DEPENDENT VARIABLE: AVERAGE PERCENTILE SCORE ON SAT

Grade	Actual test data		Actual and imputed test data	
	Coefficient on small class dum.	Sample size	Coefficient on small class dum.	Sample size
K	5.32 (.76)	5900	5.32 (.76)	5900
1	6.95 (.74)	6632	6.30 (.68)	8328
2	5.59 (.76)	6282	5.64 (.65)	9773
3	5.58 (.79)	6339	5.49 (.63)	10919

Estimates of reduced-form models are presented. Each regression includes the following explanatory variables: a dummy variable indicating initial assignment to a small class; a dummy variable indicating initial assignment to a regular/aide class, unrestricted school effects; a dummy variable for student gender; and a dummy variable for student race. The reported coefficient on small class dummy is relative to regular classes. Standard errors are in parentheses.

As you can see in Figure 10, there is nothing in the analysis that suggests bias has crept in because of attrition.

Whereas attrition is a problem of units *leaving* the experiment altogether, there's also a problem in which students *switch* between treatment status. A contemporary example of this was in the antiretroviral treatment experiments for HIV in the 1980s. These experiments were often contaminated by the fact that secondary markets for the experimental treatments formed in which control groups purchased the treatment. Given the high stakes of life and death associated with HIV at the time, this is understandable, but scientifically, this switching of treatment status contaminated the experiment making estimates of the *ATE* biased. Krueger writes that in this educational context

Figure 10: Regression of first grade percentile scores onto treatments for K-3 with imputed test scores for all post-kindergarten ages [Krueger, 1999].

"It is virtually impossible to prevent some students from switching between class types over time." (Krueger [1999] p. 506)

To illustrate this problem of switching, Krueger created a helpful *transition matrix* which is shown in Figure 11. If students in the second grade had remained in their first grade classes, then the off-diagonal elements of this transition matrix would be zero. But because they are not zero, it means there was some switching. Of the 1,482 first graders assigned to small classrooms, 1,435 remained in small classes, and 23 and 24 switched into the other kinds. If students with stronger expected academic potential were more likely to move into small classes, then these transitions would bias the simple comparison of outcomes upwards, making small class sizes appear to be more effective than they really are.

B. First grade to second grade

First grade	Second grade			
	Small	Regular	Reg/aide	All
Small	1435	23	24	1482
Regular	152	1498	202	1852
Aide	40	115	1560	1715
All	1627	1636	1786	5049

Figure 11: Switching of students into and out of the treatment arms between first and second grade [Krueger, 1999].

What can be done to address switching under random assignment? Well, one thing that *could've* been done is to make it very difficult. One of the things that characterizes the modern random experiment is designing the experiment in such a way that makes switching very hard. But this may be practically infeasible, so a second best solution is to regress the student's outcome against the original randomized kindergarten class size, as opposed to the actual class size – a kind of reduced form instrumental variables approach.⁶³ If a student had been randomly assigned to a small class, but switched to a regular class in the first grade, we would regress scores on the original assignment since the original assignment satisfies the independence assumption. So long as this original assignment is highly correlated with the first through third grade class size (even if not perfectly correlated), then this regression is informative of the effect of class size on test scores. This is what makes the original assignment a good instrumental variable – because it is highly correlated with subsequent class size, even with switching.

In this approach, kindergarten is the same for both the OLS and reduced form IV approach, because the randomization assignment and the actual classroom enrollment are the same in kindergarten. But from grade 1 onwards, OLS and reduced form IV differ because of the switching.

⁶³ We will discuss this in more detail when we cover instrumental variables (IV). But it is necessary to at least cover bits of IV now since this is a common second best solution in a randomized experiment when switching occurs.

Explanatory variable	OLS: actual class size				Reduced form: initial class size			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
B. First grade								
Small class	8.57 (1.97)	8.43 (1.21)	7.91 (1.17)	7.40 (1.18)	7.54 (1.76)	7.17 (1.14)	6.79 (1.10)	6.37 (1.11)
Regular/aide class	3.44 (2.05)	2.22 (1.00)	2.23 (0.98)	1.78 (0.98)	1.92 (1.12)	1.69 (0.80)	1.64 (0.76)	1.48 (0.76)
White/Asian (1 = yes)	—	—	6.97 (1.18)	6.97 (1.19)	—	—	6.86 (1.18)	6.85 (1.18)
Girl (1 = yes)	—	—	3.80 (.56)	3.85 (.56)	—	—	3.76 (.56)	3.82 (.56)
Free lunch (1 = yes)	—	—	-13.49 (.87)	-13.61 (.87)	—	—	-13.65 (.88)	-13.77 (.87)
White teacher	—	—	—	-4.28 (1.96)	—	—	—	-4.40 (1.97)
Male teacher	—	—	—	11.82 (3.33)	—	—	—	13.06 (3.38)
Teacher experience	—	—	—	.05 (0.06)	—	—	—	.06 (.06)
Master's degree	—	—	—	.48 (1.07)	—	—	—	.63 (1.09)
School fixed effects	No .02	Yes .24	Yes .30	Yes .30	No .01	Yes .23	Yes .29	Yes .30

Figure 12: IV reduced form approach compared to the OLS approach [Krueger, 1999].

Figure 12 shows eight regressions – four per approach, where each four is like the ones shown in previous figures. Briefly, just notice that while the two regressions yield different coefficients, their magnitudes and precision are fairly similar. The reduced form IV approach yields coefficients that are about 1 percentage point lower on average than what he got with OLS.

Some other problems worth mentioning when it comes to randomized experiments. First, there could be heterogeneous treatment effects. In other words, perhaps δ_i differs across i students. If this is the case, then $ATT \neq ATU$ though in large enough samples, and under the independence assumption, this difference should be negligible.

Now we do our own analysis. Go back into Stata and type:

```
. reg tscorek sck rak
```

Krueger standardized the test scores into percentiles, but I will keep the data in its raw form simplicity. This means the results will be dissimilar to what is shown in his Figure 6.

In conclusion, we have done a few things in this chapter. We've introduced the Rubin causal model by introducing its powerful potential outcomes notation. We showed that the simple difference in mean outcomes was equal to the sum of the average treatment effect, a term called the selection bias, and a term called the weighted heterogeneous treatment effect bias. Thus the simple difference in mean outcomes estimator is biased unless those second and third terms

Dependent variable	Total kindergarten score (unscaled)
Small class	13.90 (2.41)
Regular/aide class	0.314 (2.310)

Table 13: Krueger regressions [Krueger, 1999].

zero out. One situation in which they zero out is under *independence* of the treatment, which is when the treatment has been assigned independently of the potential outcomes. When does independence occur? The most commonly confronted situation where it would occur is under physical randomization of the treatment to the units. Because physical randomization assigns the treatment for reasons that are *independent* of the potential outcomes, the selection bias zeroes out as does the heterogeneous treatment effect bias. We now move to discuss a second situation where the two terms zero out: *conditional* independence.

Matching and subclassification

One of the main things I wanted us to learn from the chapter on directed acylic graphical models is the idea of the backdoor criterion. Specifically, if there exists a conditioning strategy that will satisfy the backdoor criterion, then you can use that strategy to identify your causal effect of interest. We now discuss three different kinds of conditioning strategies. They are subclassification, exact matching, and approximate matching. I will discuss each of them now.

Subclassification

Subclassification is a method of satisfying the backdoor criterion by weighting differences in means by strata-specific weights. These strata-specific weights will, in turn, adjust the differences in means so that their distribution by strata is the same as that of the counterfactual's strata. This method implicitly achieves distributional *balance* between the treatment and control in terms of that known, observable confounder. This method was created by statisticians like [Cochran \[1968\]](#) when trying to analyze the causal effect of smoking on lung cancer, and while the methods today have moved beyond it, we include it because some of the techniques implicit in subclassification are present throughout the rest of the book.⁶⁴

One of the concepts that will thread through this chapter is the concept of the conditional independence assumption, or CIA. In the previous example with the STAR test, we said that the experimenters had assigned the small classes to students *conditionally* randomly. That is, conditional on a given school, α_s , the experimenters randomly assigned the treatment across teachers and students. This technically meant that treatment was independent of potential outcomes *for any given school*. This is a kind of independence assumption, but it's one that must incorporate the conditional element to the independent assignment of the treatment. assumption is written as

$$(Y^1, Y^0) \perp\!\!\!\perp D | X$$

where again $\perp\!\!\!\perp$ is the notation for statistical independence and X

⁶⁴ To my knowledge, [Cochran \[1968\]](#) is the seminal paper on subclassification.

is the variable we are conditioning on. What this means is that the expected values of Y^1 and Y^0 are equal for treatment and control group *for each value of X*. Written out this means:

$$\begin{aligned} E[Y^1|D = 1, X] &= E[Y^1|D = 0, X] \\ E[Y^0|D = 1, X] &= E[Y^0|D = 0, X] \end{aligned}$$

Put into words, the expected value of each potential outcome is equal for the treatment group and the control group, once we condition on some X variable. If CIA can be credibly assumed, then it necessarily means you have selected a conditioning strategy that satisfies the backdoor criterion. They are equivalent concepts as far as we are concerned.

An example of this would mean that for the ninth school in our sample, $\alpha_9 = 1$, the expected potential outcomes are the same for small and large classes, and so on. When treatment is conditional on observable variables, such that the CIA is satisfied, we say that the situation is one of *selection on observables*. If one does not directly address the problem of selection on observables, estimates of the treatment effect will be biased. But this is remedied if the observable variable is conditioned on. The variable X can be thought of as an $n \times k$ matrix of covariates which satisfy the CIA as a whole.

I always find it helpful to understand as much history of thought behind econometric estimators, so let's do that here. One of the public health problems of the mid to late 20th century was the problem of rising lung cancer. From 1860 to 1950, the incidence of lung cancer in cadavers grew from 0% of all autopsies to as high as 7% (Figure 13). The incidence appeared to be growing at an increasing rate. The mortality rate per 100,000 from cancer of the lungs in males reached 80-100 per 100,000 by 1980 in the UK, Canada, England and Wales. From 1860 to 1950, the incidence of lung cancer in cadavers grew from 0% of all autopsies to as high as 7% (Figure 13). The incidence appeared to be growing at an increasing rate. The mortality rate per 100,000 from cancer of the lungs in males reached 80-100 per 100,000 by 1980 in the UK, Canada, England and Wales.

Several studies were found to show that the odds of lung cancer was directly related to the amount of cigarettes the person smoked a day. Figure 14 shows that the relationship between daily smoking and lung cancer in males was monotonic in the number of cigarettes the male smoked per day. Smoking, it came to be believed, was *causing* lung cancer. But some statisticians believed that scientists couldn't draw that conclusion because it was possible that smoking was not independent of health outcomes. Specifically, perhaps the people who smoked cigarettes differed from one another in ways that

Figure 1
Lung Cancer at Autopsy: Combined Results from 18 Studies

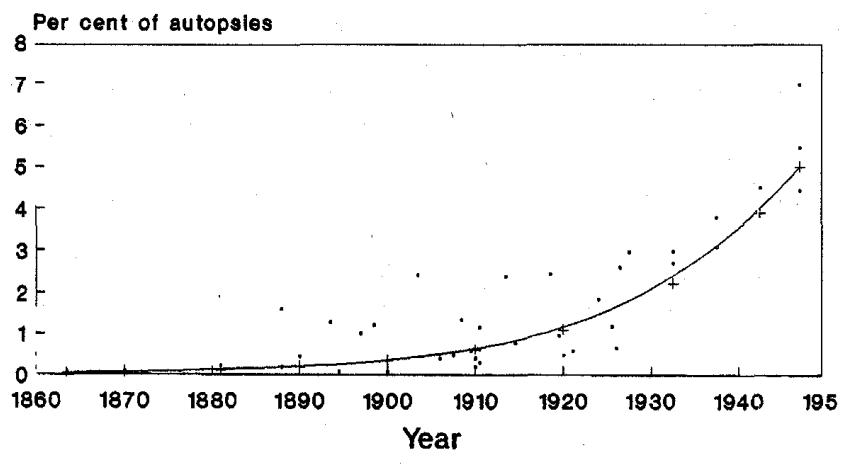
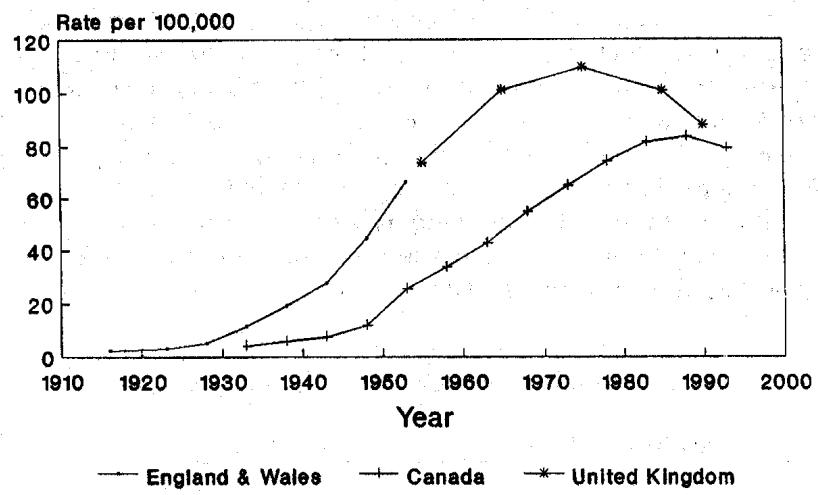


Figure 13: Lung cancer at autopsy trends

Figure 2(a)
Mortality from Cancer of the Lung in Males



were directly related to the incidence of lung cancer.

This was a classic “correlation does not necessarily mean causation” kind of problem. Smoking was clearly correlated with lung cancer, but does that necessarily mean that smoking *caused* lung cancer? Thinking about the simple difference in means notation, we know that a comparison of smokers and non-smokers will be biased in observational data if the independence assumption does not hold. And because smoking is endogenous, meaning people choose to smoke, it’s entirely possible that smokers differed from the non-smokers in ways that were directly related to the incidence of lung cancer. Criticisms at the time came from such prominent statisticians as Joseph Berkson, Jerzy Neyman and Ronald Fisher. Their reasons were as follows. First, it was believed that the correlation was spurious because of a biased selection of subjects. The samples were non-random in other words. Functional form complaints were also common. This had to do with people’s use of risk ratios and odds ratios. The association, they argued, was sensitive to those kinds of functional form choices.

Probably most damning, though, was the hypothesis that there existed an unobservable genetic element that both caused people to smoke and which, independently, caused people to develop lung cancer [Pearl, 2009]. This confounder meant that smokers and non-smokers differed from one another in ways that were directly related to their potential outcomes, and thus independence did not hold. Other studies showed that cigarette smokers and non-smokers differed on observables. For instance, smokers were more extroverted than non-smokers, as well as differed in age, income, education, and so on.

Other criticisms included that the magnitudes relating smoking and lung cancer were considered implausibly large. And again, the ever present criticism of observational studies, there did not exist any experimental evidence that could incriminate smoking as a cause of lung cancer.

But think about the hurdle that that last criticism actually creates. Imagine the hypothetical experiment: a large sample of people, with diverse potential outcomes, are assigned to a treatment group (smoker) and control (non-smoker). These people must be dosed with their corresponding treatments long enough for us to observe lung cancer develop – so presumably years of heavy smoking. How could anyone ever run an experiment like that? To describe it is to confront the impossibility of running such a randomized experiment. But how then do we answer the causal question without independence (i.e., randomization)?

It is too easy for us to criticize Fisher and company for their stance

Figure 4
Smoking and Lung Cancer Case-control Studies

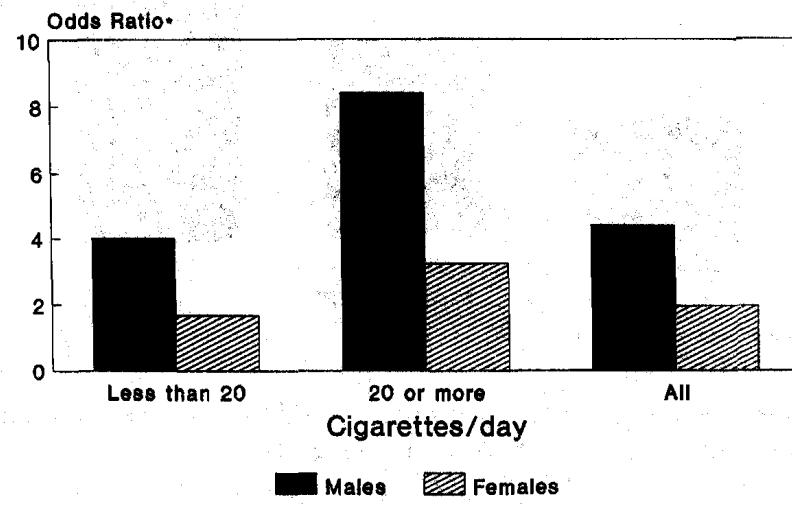
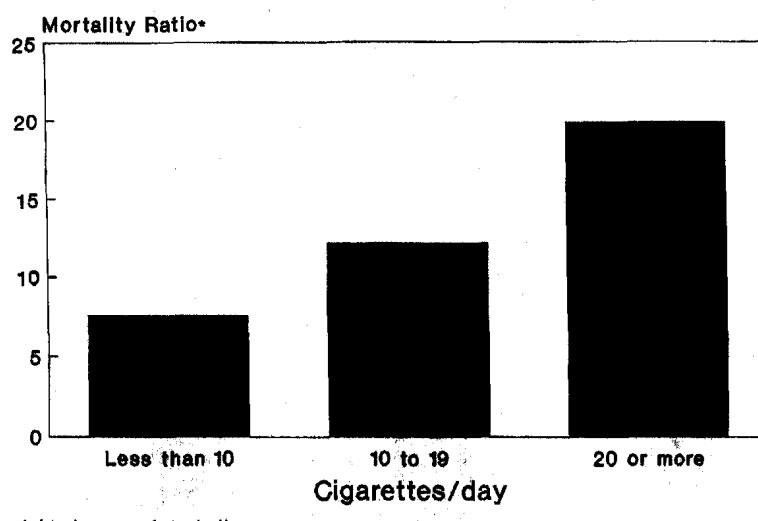


Figure 14: Smoking and Lung Cancer

Figure 5
Smoking and Lung cancer Cohort Studies in Males



on smoking as a causal link to lung cancer because that causal link is now universally accepted as scientific fact. But remember, it was not always. And the correlation/causation point is a damning one. Fisher's arguments, it turns out, was based on sound science.⁶⁵ Yet, we now know that in fact the epidemiologists were right. Hooke [1983] wrote:

"the [epidemiologists] turned out to be right, but only because bad logic does not necessarily lead to wrong conclusions."

To motivate what we're doing in subclassification, let's work with Cochran [1968], which was a study trying to address strange patterns in smoking data by adjusting for a confounder.⁶⁶ Cochran lays out mortality rates by country and smoking type (Table 14).

Smoking group	Canada	British	US
Non-smokers	20.2	11.3	13.5
Cigarettes	20.5	14.1	13.5
Cigars/pipes	35.5	20.7	17.4

As you can see, the highest death rate among Canadians is the cigar and pipe smokers, which is considerably higher than that of non-smokers or cigarettes. Similar patterns show up in both countries, though smaller in magnitude than what we see in Canada.

This table suggests that pipes and cigar smoking are more dangerous than cigarette smoking which, to a modern reader, sounds ridiculous. The reason it sounds ridiculous is because cigar and pipe smokers often do not inhale, and therefore there is less tar that accumulates in the lungs than with cigarettes. And insofar as it's the tar that causes lung cancer, it stands to reason that we should see higher mortality rates among cigarette smokers.

But, recall the independence assumption. Do we really believe that:

$$\begin{aligned} E[Y^1 | \text{Cigarette}] &= E[Y^1 | \text{Pipe}] = E[Y^1 | \text{Cigar}] \\ E[Y^0 | \text{Cigarette}] &= E[Y^0 | \text{Pipe}] = E[Y^0 | \text{Cigar}] \end{aligned}$$

Is it the case that factors related to these three states of the world are truly independent to the factors that determine death rates? One way to check this is to see if the three groups are *balanced* on pre-treatment covariates. If the means of the covariates are the same for each group, then we say those covariates are balanced and the two groups are *exchangeable* with respect to those covariates.

One variable that appears to matter is the age of the person. Older people were more likely at this time to smoke cigars and pipes, and

⁶⁵ But, it is probably not a coincidence that Roland Fisher, the harshest critic of the epidemiological theory that smoking caused lung cancer, was himself a chain smoker. When he died of lung cancer, he was the highest paid expert witness for the tobacco industry in history.

⁶⁶ I first learned of this paper from Alberto Abadie in a lecture he gave at the Northwestern Causal Inference workshop.

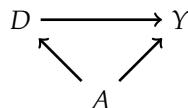
Table 14: Death rates per 1,000 person-years [Cochran, 1968]

without stating the obvious, older people were more likely to die. In Table 15 we can see the mean ages of the different groups.

Smoking group	Canada	British	US
Non-smokers	54.9	49.1	57.0
Cigarettes	50.5	49.8	53.2
Cigars/pipes	65.9	55.7	59.7

Table 15: Mean ages, years [Cochran, 1968].

The high means for cigar and pip smokers are probably not terribly surprising to most of you. Cigar and pipe smokers are typically older than cigarette smokers, or at least were in 1968 when this was written. And since older people die at a higher rate (for reasons other than just smoking cigars), maybe the higher death rate for cigar smokers is because they're older on average. Furthermore, maybe by the same logic the reason that cigarette smoking has such a low mortality rate is because cigarette smokers are younger on average. Note, using DAG notation, this simply means that we have the following DAG:



where D is smoking, Y is mortality, and A is age of the smoker. Insofar as CIA is violated, then we have a backdoor path that is open, which also means in the traditional pedagogy that we have omitted variable bias. But however we want to describe it, the common thing it will mean is that the distribution of age for each group will be different – which is what I mean by *covariate imbalance*. My first strategy for addressing this problem of covariate balance is by *condition* on the key variable, which in turn will *balance* the treatment and control groups with respect to this variable.⁶⁷

So how do we exactly close this backdoor path using subclassification? We calculate the mortality rate for some treatment group (cigarette smokers) by some strata (here, that is age). Next, we then weight the mortality rate for the treatment group by a strata (age)-specific weight that corresponds to the control group. This gives us the age adjusted mortality rate for the treatment group. Let's explain with an example by looking at Table 16. Assume that age is the only relevant confounder between cigarette smoking and mortality. Our first step is to divide age into strata: say 20-40, 41-70, and 71 and older.

What is the average death rate for pipe smokers without subclassification? It is the weighted average of the mortality rate column

⁶⁷ This issue of covariate balance runs throughout nearly every identification strategy that we will discuss, in some way or another.

Table 16: Subclassification example.

	Death rates		Number of Pipe/cigar-smokers
	Cigarette-smokers	Cigarette-smokers	
Age 20-40	20	65	10
Age 41-70	40	25	25
Age ≥ 71	60	10	65
Total		100	100

where each weight is equal to $\frac{N_t}{N}$ and N_t and N are the number of people in each group and the total number of people, respectively. Here that would be $20 \times \frac{65}{100} + 40 \times \frac{25}{100} + 60 \times \frac{10}{100} = 29$. That is, the mortality rate of smokers in the population is 29 per 100,000.

But notice that the age distribution of cigarette smokers is the exact opposite (by construction) of pipe and cigar smokers. Thus the age distribution is imbalanced. Subclassification simply adjusts the mortality rate for cigarette smokers so that it has the same age distribution of the comparison group. In other words, we would multiply each age-specific mortality rate by the proportion of individuals in that age strata for the comparison group. That would be

$$20 \times \frac{10}{100} + 40 \times \frac{25}{100} + 60 \times \frac{65}{100} = 51$$

That is, when we adjust for the age distribution, the age-adjusted mortality rate for cigarette smokers (were they to have the same age distribution as pipe and cigar smokers) would be 51 per 100,000 – almost twice as large as we got taking a simple naive calculation unadjusted for the age confounder.

Cochran uses a version of this subclassification method in his paper and recalculates the mortality rates for the three countries and the three smoking groups. See Table 17. As can be seen, once we adjust for the age distribution, cigarette smokers have the highest death rates among any group.

Smoking group	Canada	UK	US
Non-smokers	20.2	11.3	13.5
Cigarettes	29.5	14.8	21.2
Cigars/pipes	19.8	11.0	13.7

Table 17: Adjusted mortality rates using 3 age groups [Cochran, 1968].

Which variables should be used for adjustments? This kind of adjustment raises a question – which variable should we use for adjustment. First, recall what we've emphasized repeatedly. Both the backdoor criterion and CIA tell us precisely what we need to do. We

need to choose a variable that once we condition on it, all backdoor paths are closed and therefore the CIA is met. We call such a variable the *covariate*. A covariate is usually a random variable assigned to the individual units prior to treatment. It is predetermined and therefore exogenous. It is not a collider, nor is it endogenous to the outcome itself (i.e., no conditioning on the dependent variable). A variable is exogenous with respect to D if the value of X does not depend on the value of D . Oftentimes, though not always and not necessarily, this variable will be time invariant, such as race.

Why shouldn't we include in our adjustment some descendent of the outcome variable itself? We saw this problem in our first collider example from the DAG chapter. Conditioning on a variable that is a descendent of the outcome variable can introduce collider bias, and it is not easy to know *ex ante* just what kind of bias this will introduce.

Thus, when trying to adjust for a confounder using subclassification, let the DAG help you choose which variables to include in the conditioning strategy. Your goal ultimately is to satisfy the backdoor criterion, and if you do, then the CIA will hold in your data.

Identifying assumptions Let me now formalize what we've learned. In order to estimate a causal effect when there is a confounder, we need (1) CIA and (2) the probability of treatment to be between 0 and 1 for each strata. More formally,

1. $(Y^1, Y^0) \perp\!\!\!\perp D|X$ (conditional independence)
2. $0 < Pr(D = 1|X) < 1$ with probability one (common support)

These two assumptions yield the following identity

$$\begin{aligned} E[Y^1 - Y^0|X] &= E[Y^1 - Y^0|X, D = 1] \\ &= E[Y^1|X, D = 1] - E[Y^0|X, D = 0] \\ &= E[Y|X, D = 1] - E[Y|X, D = 0] \end{aligned}$$

where each value of Y is determined by the switching equation.

Given common support, we get the following estimator:

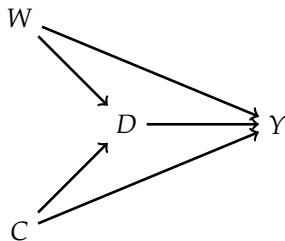
$$\widehat{\delta}_{ATE} = \int (E[Y|X, D = 1] - E[Y|X, D = 0]) dPr(X)$$

These assumptions are necessary to identify the ATE, but fewer assumptions are needed. They are that D is conditionally independent of Y^0 , and that there exists some units in the control group for each treatment strata. Note, the reason for the common support assumption is because we are weighting the data; without common support, we cannot calculate the relevant weights.

Subclassification exercise: Titanic dataset For what we are going to do next, I find it useful to move into actual data. We will use a dataset to conduct subclassification which I hope you find interesting. As everyone knows, the Titanic ocean cruiser hit an iceberg and sank on its maiden voyage. A little over 700 passengers and crew survived out of the 2200 total. It was a horrible disaster. Say that we are curious as to whether or not seated in first class, were you more likely to survive. To answer this, as always, we need two things: data and assumptions.

. scuse titanic, clear

Our question as to whether first class seating increased the probability of survival is confounded by the oceanic norms during disasters. Women and children should be escorted to the lifeboats before the men in the event of a disaster requiring exiting the ship. If more women and children were in first class, then maybe first class is simply picking up the effect of that social norm, rather than the effect of class and wealth on survival. Perhaps a DAG might help us here, as a DAG can help us outline the sufficient conditions for identifying the causal effect of first class on survival.



Now before we commence, let's review what it means. This says that being a female made you more likely to be in first class, but also made you more likely to survive because lifeboats were more likely to be allocated to women. Furthermore, being a child made you more likely to be in first class and made you more likely to survive. Finally, there are no other confounders, observed or unobserved.⁶⁸

Here we have one direct path (the causal effect) between first class (D) and survival (Y) and that's $D \rightarrow Y$. But, we have two backdoor paths. For instance, we have the $D \leftarrow C \rightarrow Y$ backdoor path and we have the $D \leftarrow W \rightarrow Y$ backdoor path. But fortunately, we have data that includes both age and gender, so it is possible to close each backdoor path and therefore satisfy the backdoor criterion. We will use subclassification to do that.

But, before we use subclassification to achieve the backdoor criterion, let's calculate a naive simple difference in outcomes (SDO) which is just

$$E[Y|D = 1] - E[Y|D = 0]$$

⁶⁸ I'm sure you can think of others, though, in which case this DAG is misleading.

for the sample.

```

. gen female=(sex==0)
. label variable female "Female"
. gen male=(sex==1)
. label variable male "Male"
. gen s=1 if (female==1 & age==1)
. replace s=2 if (female==1 & age==0)
. replace s=3 if (female==0 & age==1)
. replace s=4 if (female==0 & age==0)
. gen d=1 if class==1
. replace d=0 if class!=1
. summarize survived if d==1
. gen ey1=r(mean)
. summarize survived if d==0
. gen ey0=r(mean)
. gen sdo=ey1-ey0
. su sdo
* SDO says that being in first class raised the probability of survival by 35.4%

```

When you run this code, you'll find that the people in first class were 35.4% more likely to survive than people in any other group of passengers including the crew. But note, this does not take into account the confounders of age and gender. So next we use subclassification weighting to control for these confounders. Here's the steps that that will entail:

1. Stratify the data into four groups: young males, young females, old males, old females
2. Calculate the difference in survival probabilities for each group
3. Calculate the number of people in the non-first class groups and divide by the total number of non-first class population. These are our strata specific weights.
4. Calculate the weighted average survival rate using the strata weights.

Let's do this in Stata, which hopefully will make these steps more concrete.

```
. cap n drop ey1 ey0
. su survived if s==1 & d==1
. gen ey11=r(mean)
. label variable ey11 "Average survival for male child in treatment"
. su survived if s==1 & d==0
. gen ey10=r(mean)
. label variable ey10 "Average survival for male child in control"
. gen diff1=ey11-ey10
. label variable diff1 "Difference in survival for male children"
. su survived if s==2 & d==1
. gen ey21=r(mean)
. su survived if s==2 & d==0
. gen ey20=r(mean)
. gen diff2=ey21-ey20
. su survived if s==3 & d==1
. gen ey31=r(mean)
. su survived if s==3 & d==0
. gen ey30=r(mean)
. gen diff3=ey31-ey30
. su survived if s==4 & d==1
. gen ey41=r(mean)
. su survived if s==4 & d==0
. gen ey40=r(mean)
. gen diff4=ey41-ey40
. count if s==1 & d==0
. count if s==2 & d==0
. count if s==3 & d==0
. count if s==4 & d==0
. count
. gen wt1=425/2201
. gen wt2=45/2201
. gen wt3=1667/2201
. gen wt4=64/2201
. gen wate=diff1*wt1 + diff2*wt2 + diff3*wt3 + diff4*wt4
. sum wate sdo
```

Here we find that once we condition on the confounders, gender and age, we find a much lower probability of survival associated with first class (though frankly, still large). The weighted ATE is 16.1% vs the SDO which is 35.4%.

Curse of dimensionality Here we've been assuming two covariates each of which has two possible set of values. But this was for convenience. Our Titanic dataset, for instance, only came to us with two possible values for age – child and adult. But what if it had come to us with multiple values for age, like specific age? Then once we condition on individual age and gender, it's entirely likely that we will not have the information necessary to calculate differences within strata, and therefore be unable to calculate the strata-specific weights that we need for subclassification.

For this next part, let's assume that we have precise data on Titanic survivor ages. But because this will get incredibly laborious, let's just focus on a few of them.

Age and Gender	Survival Prob			Number of	
	1st Class	Controls	Diff.	1st Class	Controls
Male 11-yo	1.0	0	1	1	2
Male 12-yo	–	1	–	0	1
Male 13-yo	1.0	0	1	1	2
Male 14-yo	–	0.25	–	0	4
...					

Table 18: Subclassification example of Titanic survival for large K

Here we see an example of the common support assumption being violated. The common support assumption requires that for each strata, there exist observations in both the treatment and control group, but as you can see, there are not any 12 year old male passengers in first class. Nor are there any 14-year old male passengers in first class. And if we were to do this for every age \times gender combination, we would find that this problem was quite common. Thus we cannot calculate the ATE.

But, let's say that the problem was always on the treatment group, not the control group. That is, let's assume that there is always *someone* in the control group for a given gender \times age combination, but there isn't always for the treatment group. Then we can calculate the ATT. Because as you see in this table, for those two strata, 11 and 13 year olds, there is both treatment and control group values for the calculation. So long as there exists controls for a given treatment strata, we can calculate the ATT. The equation to do so can be compactly

written as:

$$\hat{\delta}_{ATT} = \sum_{k=1}^K (\bar{Y}^{1,k} - \bar{Y}^{0,k}) \times \left(\frac{N_T^k}{N_T} \right)$$

Plugging in values for those summations, we get

We've seen now a problem that arises with subclassification – in a finite sample, subclassification becomes less feasible as the number of covariates grows because as K grows, the data becomes sparse. We will at some point be missing values, in other words, for those K categories. Imagine if we tried to add a third strata, say race (black and white). Then we'd have two age categories, two gender categories and two race categories, giving us eight possibilities. In this small sample, we probably will end up with many cells having missing information. This is called the *curse of dimensionality*. If sparseness occurs, it means many cells may contain either only treatment units, or only control units, but not both. If that happens, we can't use subclassification, because we do not have common support. And therefore we are left searching for an alternative method to satisfy the backdoor criterion.

Exact matching

Subclassification uses the difference between treatment and control group units, and achieves covariate balance by using the K probability weights to weight the averages. It's a simple method, but it has the aforementioned problem of the "curse of dimensionality". And probably, that's going to be an issue practically in any research you undertake because it may not be merely one variable you're worried about, but several. In which case, you'll already be running into the curse. But the thing that we emphasize here is that the subclassification method is using the raw data, but weighting it so as to achieve balance. We are weighting the differences, and then summing over those weighted differences.

But there's alternative approaches. For instance, what if we estimated $\hat{\delta}_{ATT}$ by *imputing* the missing potential outcomes by conditioning on the confounding, observed covariate? Specifically, what if we filled in the missing potential outcome for each treatment unit using a control group unit that was "closest" to the treatment group unit for some X confounder. This would give us estimates of all the counterfactuals from which we could simply take the average over the differences. As we will show, this will also achieve covariate balance. This method is called *matching*.

There are two broad types of matching that we will consider: exact

matching and approximate matching. We will first start by describing exact matching. Much of what I am going to be discussing is based on Abadie and Imbens [2006].⁶⁹

A simple matching estimator is the following:

$$\hat{\delta}_{ATT} = \frac{1}{N_T} \sum_{D_i=1} (Y_i - Y_{j(i)})$$

where $Y_{j(i)}$ is the j^{th} unit matched to the i^{th} unit based on the j^{th} being “closest to” the i^{th} unit for some X covariate. For instance, let’s say that a unit in the treatment group has a covariate with value 2 and we find another unit in the control group (exactly one unit) with a covariate value of 2. Then we will impute the treatment unit’s missing counterfactual with the matched unit’s, and take a difference.

But, what if there’s more than one variable “closest to” the i^{th} unit? For instance, say that the same i^{th} unit has a covariate value of 2 and we find two j units with a value of 2. What can we then do? Well, one option is to simply take the average of those two units’ Y outcome value. What if we find 3? What if we find 4, and so on? However many matches M that we find, we would assign the average outcome ($\frac{1}{M}$) as the counterfactual for the treatment group unit.

Notationally, we can describe this estimator as

$$\hat{\delta}_{ATT} = \frac{1}{N_T} \sum_{D_i=1} \left(Y_i - \left[\frac{1}{M} \sum_{m=1}^M Y_{j_m(1)} \right] \right)$$

This really isn’t too different of an estimator from the one before it; the difference is the $\frac{1}{M}$ which is the averaging over closest matches that we were talking about. This approach works well when we can find a number of good matches for each treatment group unit. We usually define M to be small, like $M = 2$. If there are more than 2, then we may simply randomly select two units to average outcomes over.⁷⁰

Those were all average treatment effects on the treatment group estimators. You can tell that these are $\hat{\delta}_{ATT}$ estimators because of the summing over the treatment group.⁷¹ But we can also estimate the ATE . But note, when estimating the ATE , we are filling in both (a) missing control group units like before and (b) missing treatment group units. If observation i is treated, in other words, then we need to fill in the missing Y_i^0 using the control matches, and if the observation i is a control group unit, then we need to fill in the missing Y_i^1 using the treatment group matches. The estimator is below. It looks scarier than it really is. It’s actually a very compact, nicely written out estimator equation.

⁶⁹ I first learned about this form of matching from lectures by Alberto Abadie at the Northwestern Causal Inference workshop – a workshop that I highly recommend.

⁷⁰ Note that all of these approaches require some programming, as they’re algorithms.

⁷¹ Notice the $D_i = 1$ in the subscript of the summation operator.

$$\hat{\delta}_{ATE} = \frac{1}{N} \sum_{i=1}^N (2D_i - 1) \left[Y_i - \left(\frac{1}{M} \sum_{m=1}^M Y_{j_m(i)} \right) \right]$$

The $2D_i - 1$ is the nice little trick. When $D_i = 1$ then that leading term becomes a 1.⁷² And when $D_i = 0$, then that leading term becomes a negative 1, and the outcomes reverse order so that the treatment observation can be imputed. Nice little mathematical form!

⁷² $2 \times 1 - 1 = 1$.

Let's see this work in action by working with an example. Table 19 shows two samples: a list of participants in a job trainings program and a list of non-participants, or non-trainees. The left-hand-side group is the treatment group and the right-hand-side group is the control group. The matching algorithm that we defined earlier will create a third group called the *matched sample* consisting of each treatment group unit's matched counterfactual. Here we will match on the age of the participant.

Trainees			Non-Trainees		
Unit	Age	Earnings	Unit	Age	Earnings
1	18	9500	1	20	8500
2	29	12250	2	27	10075
3	24	11000	3	21	8725
4	27	11750	4	39	12775
5	33	13250	5	38	12550
6	22	10500	6	29	10525
7	19	9750	7	39	12775
8	20	10000	8	33	11425
9	21	10250	9	24	9400
10	30	12500	10	30	10750
			11	33	11425
			12	36	12100
			13	22	8950
			14	18	8050
			15	43	13675
			16	39	12775
			17	19	8275
			18	30	9000
			19	51	15475
			20	48	14800
Mean	24.3	\$11,075		31.95	\$11,101.25

Table 19: Training example with exact matching

Before we do this, though, I want to show you how the ages of the trainees differs on average with the ages of the non-trainees.

We can see that in Table 19 – the average age of the participants is 24.3 and the average age of the non-participants is 31.95. Thus the people in the control group are older, and since wages typically rise with age, we may suspect that part of the reason that their average earnings is higher (\$11,075 vs. \$11,101) is because the control group is older. We say that the two groups are not *exchangeable* because the covariate is not *balanced*. Let's look at the age distribution ourselves to see. To illustrate this, let's download the data first. We will create two histograms – the distribution of age for treatment and non-trainee group – as well as summarize earnings for each group. That information is also displayed in Figure 15.

```
. scuse training_example, clear
. histogram age_treat, bin(10) frequency
. histogram age_controls, bin(10) frequency
. su age_treat age_controls
. su earnings_treat earnings_control
```

As you can see from Figure 15, these two populations not only have different means (Table 19), but the entire distribution of age across the samples is different. So let's use our matching algorithm and create the missing counterfactuals for each treatment group unit. This method, since it only imputes the missing units for each treatment unit, will yield an estimate of the $\hat{\delta}_{ATT}$.

Now let's move to creating the matched sample. As this is exact matching, the distance traveled to the nearest neighbor will be zero integers. This won't always be the case, but note that as the control group sample size grows, the likelihood we find a unit with the same covariate value as one in the treatment group grows. I've created a dataset like this. The first treatment unit has an age of 18. Searching down through the non-trainees, we find exactly one person with an age of 18 and that's unit 14. So we move the age and earnings information to the new matched sample columns.

We continue doing that for all units, always moving the control group unit with the closest value on X to fill in the missing counterfactual for each treatment unit. If we run into a situation where there's more than one control group unit "close", then we simply average over them. For instance, there are two units in the non-trainees group with an age of 30, and that's 10 and 18. So we averaged their earnings and matched that average earnings to unit 10. This is filled out in Table 20.

Now we see that the mean age is the same for both groups. We can also check the overall age distribution (Figure 16). As you can

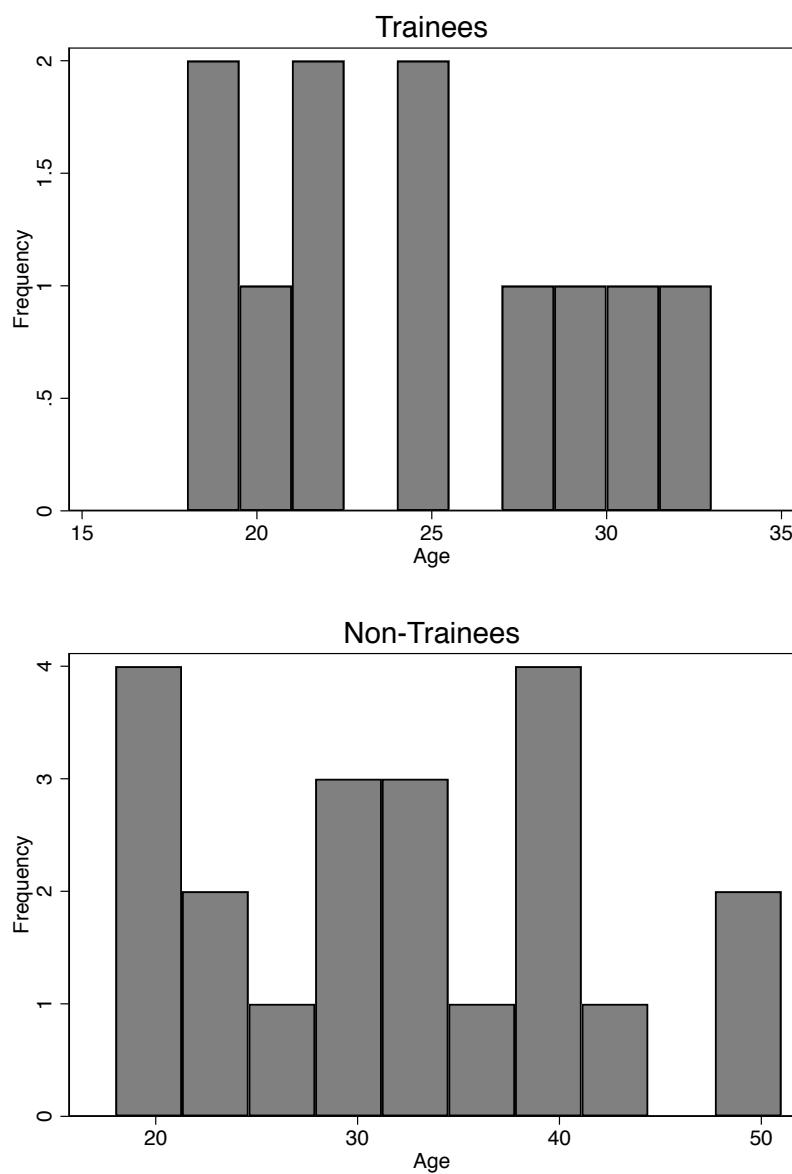


Figure 15: Covariate distribution by job trainings and control.

Table 20: Training example with exact
matching (including matched sample)

Trainees			Non-Trainees			Matched Sample		
Unit	Age	Earnings	Unit	Age	Earnings	Unit	Age	Earnings
1	18	9500	1	20	8500	14	18	8050
2	29	12250	2	27	10075	6	29	10525
3	24	11000	3	21	8725	9	24	9400
4	27	11750	4	39	12775	8	27	10075
5	33	13250	5	38	12550	11	33	11425
6	22	10500	6	29	10525	13	22	8950
7	19	9750	7	39	12775	17	19	8275
8	20	10000	8	33	11425	1	20	8500
9	21	10250	9	24	9400	3	21	8725
10	30	12500	10	30	10750	10,18	30	9875
			11	33	11425			
			12	36	12100			
			13	22	8950			
			14	18	8050			
			15	43	13675			
			16	39	12775			
			17	19	8275			
			18	30	9000			
			19	51	15475			
			20	48	14800			
Mean	24.3	\$11,075		31.95	\$11,101.25		24.3	\$9,380

see, the two groups are *exactly balanced* on age. Therefore we describe the two groups as *exchangeable*. And the difference in earnings between those in the treatment group and those in the control group is \$1,695. That is, we estimate that the causal effect of the program was \$1,695 in higher earnings.⁷³

⁷³ I included code for reproducing this information as well.

```
. scuse training_example, clear
. histogram age_treat, bin(10) frequency
. histogram age_matched, bin(10) frequency
. su age_treat age_controls
. su earnings_matched earnings_matched
```

Let's summarize what we've learned. We've been using a lot of different terms, drawn from different authors and different statistical traditions, so I'd like to map them onto one another. The two groups were different in ways that were directly related to both the treatment and the outcome itself. This means that the independence assumption was violated. Matching on X meant creating an exchangeable set of observations – the matched sample – and what characterized this matched sample was *balance*.

Approximate matching

The previous example of matching was relatively simple – find a unit or collection of units that have the same value of some covariate X and substitute their outcomes as some unit j's counterfactuals. Once you've done that, average the differences and you have an estimate of the ATE.

But what if when you had tried to find a match, you couldn't find another unit with that exact same value? Then you're in the world of a set of procedures that I'm calling approximate matching.

Nearest Neighbor Covariate Matching One of the instances where exact matching can break down is when the number of covariates, K , grows large. And when we have to match on more than one variable, but are not using the sub-classification approach, then one of the first things we confront is the concept of *distance*. What does it mean for one unit's covariate to be "close" to someone else's? Furthermore, what does it mean when there are multiple covariates with therefore measurements in multiple dimensions?

Matching on a single covariate is straightforward because distance is measured in terms of the covariate's own values. For instance, a distance in age is simply how close in years or months or days the

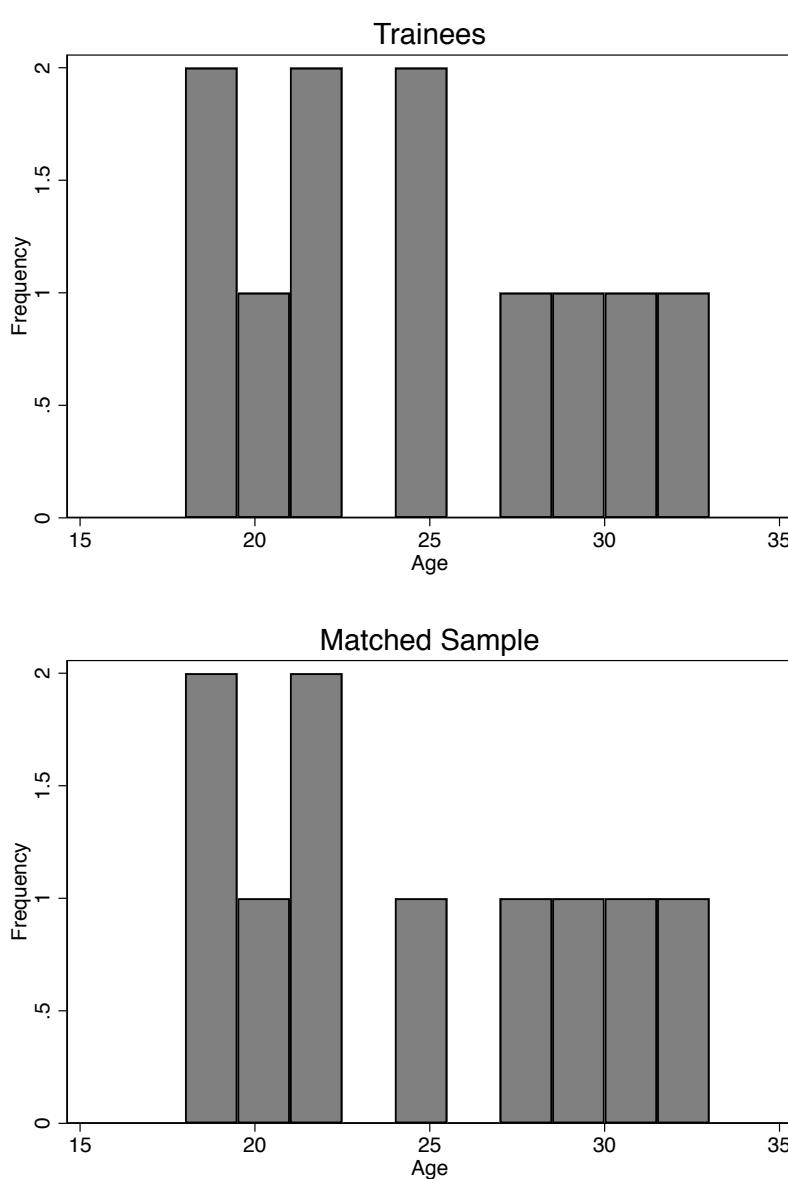


Figure 16: Covariate distribution by job trainings and matched sample.

person is to another person. But what if we have several covariates needed for matching? Say it's age and log income. A one point change in age is very different from a one point change in log income, not to mention that we are now measuring distance in two, not one, dimensions. When the number of matching covariates is more than one, we need a new definition of distance to measure closeness. We begin with the simplest measure of distance: the *Euclidean distance*:

$$\begin{aligned} ||X_i - X_j|| &= \sqrt{(X_i - X_j)'(X_i - X_j)} \\ &= \sqrt{\sum_{n=1}^k (X_{ni} - X_{nj})^2} \end{aligned}$$

The problem with this measure of distance is that the distance measure itself depends on the scale of the variables themselves. For this reason, researchers typically will use some modification of the Euclidean distance, such as the *normalized Euclidean distance*, or they'll use an alternative distance measure altogether. The normalized Euclidean distance is a commonly used distance and what makes it different is that the distance of each variable is scaled by the variable's variance. The distance is measured as:

$$||X_i - X_j|| = \sqrt{(X_i - X_j)' \hat{V}^{-1} (X_i - X_j)}$$

where

$$\hat{V}^{-1} = \begin{pmatrix} \hat{\sigma}_1^2 & 0 & \dots & 0 \\ 0 & \hat{\sigma}_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \hat{\sigma}_k^2 \end{pmatrix}$$

Notice that the normalized Euclidean distance is equal to:

$$||X_i - X_j|| = \sqrt{\sum_{n=1}^k \frac{(X_{ni} - X_{nj})^2}{\hat{\sigma}_n^2}}$$

Thus if there are changes in the scale of X , these changes also affect its variance, and so the normalized Euclidean distance does not change.

Finally, there is the *Mahalanobis* distance which like the normalized Euclidean distance measure is a scale-invariant distance metric. It is:

$$||X_i - X_j|| = \sqrt{(X_i - X_j)' \hat{\Sigma}_X^{-1} (X_i - X_j)}$$

where $\hat{\Sigma}_X$ is the sample variance-covariance matrix of X .

Basically, more than one covariate creates a lot of headaches. Not only does it create the curse of dimensionality problem, but it also makes measuring distance harder. All of this creates some challenges for finding a good match in the data. As you can see in each of these distance formulas, there are sometimes going to be matching discrepancies. Sometimes $X_i \neq X_j$. What does this mean though? It means that some unit i has been matched with some unit j on the basis of a similar covariate value of $X = x$. Maybe unit i has an age of 25, but unit j has an age of 26. Their difference is 1. Sometimes the discrepancies are small, sometimes zero, sometimes large. But, as they move away from zero, they become more problematic for our estimation and introduce bias.

How severe is this bias? First, the good news. What we know is that the matching discrepancies tend to converge to zero as the sample size increases – which is one of the main reasons that approximate matching is so data greedy. It demands a large sample size in order for the matching discrepancies to be trivially small. *But what if there are many covariates? The more covariates, the longer it takes for that convergence to zero to occur. Basically, if it's hard to find good matches with an X that has a large dimension, then you will need a lot of observations as a result. The larger the dimension, the greater likelihood of matching discrepancies, the more data you need.*

Bias correction This material is drawn from [Abadie and Imbens \[2011\]](#) which introduces bias correction techniques with matching estimators when there are matching discrepancies in finite samples. So let's begin.

Everything we're getting at is suggesting that matching is biased due to these poor matching discrepancies. So let's derive this bias. First, we write out the sample ATT estimate and then we subtract out the true ATT.

$$\hat{\delta}_{ATT} = \frac{1}{N_T} \sum_{D_i=1} (Y_i - Y_{j(i)})$$

where each i and $j(i)$ units are matched, $X_i \approx X_{j(i)}$ and $D_{j(i)} = 0$. Next we define the conditional expectation outcomes

$$\begin{aligned}\mu^0(x) &= E[Y|X = x, D = 0] = E[Y^0|X = x] \\ \mu^1(x) &= E[Y|X = x, D = 1] = E[Y^1|X = x]\end{aligned}$$

Notice, these are just the expected conditional outcome functions based on the switching equation for both control and treatment groups.

As always, we write out the observed value as a function of expected conditional outcomes and some stochastic element:

$$Y_i = \mu^{D_i}(X_i) + \varepsilon_i$$

Now rewrite the *ATT* estimator using the above μ terms:

$$\begin{aligned}\widehat{\delta}_{ATT} &= \frac{1}{N_T} \sum_{D_i=1} [(\mu^1(X_i) + \varepsilon_i) - (\mu^0(X_{j(i)}) + \varepsilon_{j(i)})] \\ &= \frac{1}{N_T} \sum_{D_i=1} (\mu^1(X_i) - \mu^0(X_{j(i)})) + \frac{1}{N_T} \sum_{D_i=1} (\varepsilon_i - \varepsilon_{j(i)})\end{aligned}$$

Notice, the first line is just the *ATT* with the stochastic element included from the previous line. And the second line rearranges it so that we get two terms: the estimated *ATT* plus the average difference in the stochastic terms for the matched sample.

Now we compare this estimator with the true value of *ATT*.

$$\begin{aligned}\widehat{\delta}_{ATT} - \delta_{ATT} &= \frac{1}{N_T} \sum_{D_i=1} (\mu^1(X_i) - \mu^0(X_{j(i)})) - \delta_{ATT} \\ &\quad + \frac{1}{N_T} \sum_{D_i=1} (\varepsilon_i - \varepsilon_{j(i)})\end{aligned}$$

which with some simple algebraic manipulation is:

$$\begin{aligned}\widehat{\delta}_{ATT} - \delta_{ATT} &= \frac{1}{N_T} \sum_{D_i=1} \left(\mu^1(X_i) - \mu^0(X_i) - \mu^0(X_{j(i)}) - \delta_{ATT} \right) \\ &\quad + \frac{1}{N_T} \sum_{D_i=1} (\varepsilon_i - \varepsilon_{j(i)}) \\ &\quad + \frac{1}{N_T} \sum_{D_i=1} \left(\mu^0(X_i) - \mu^0(X_{j(i)}) \right).\end{aligned}$$

Applying the central limit theorem and the difference, $\sqrt{N_T}(\widehat{\delta}_{ATT} - \delta_{ATT})$ converges to a normal distribution with zero mean. But, however,

$$E[\sqrt{N_T}(\widehat{\delta}_{ATT} - \delta_{ATT})] = E[\sqrt{N_T}(\mu^0(X_i) - \mu^0(X_{j(i)}))|D = 1].$$

Now consider the implications if the number of covariates is large. First, the difference between X_i and $X_{j(i)}$ converges to zero slowly. This therefore makes the difference $\mu^0(X_i) - \mu^0(X_{j(i)})$ converge to zero very slowly. Third, $E[\sqrt{N_T}(\mu^0(X_i) - \mu^0(X_{j(i)}))|D = 1]$ may not converge to zero. And fourth, $E[\sqrt{N_T}(\widehat{\delta}_{ATT} - \delta_{ATT})]$ may not converge to zero.

As you can see, the bias of the matching estimator can be severe depending on the magnitude of these matching discrepancies. However, one good piece of news is that these discrepancies are observed. We can see the degree to which each unit's matched sample has severe mismatch on the covariates themselves. Secondly, we can always

make the matching discrepancy small by using a large donor pool of untreated units to select our matches, because recall, the likelihood of finding a good match grows as a function of the sample size, and so if we are content to estimating the ATT , then increasing the size of the donor pool can buy us out of this mess. But, let's say we can't do that and the matching discrepancies are large. Then we can apply bias correction methods to minimize the size of the bias. So let's see what the bias correction method looks like. This is based on [Abadie and Imbens \[2011\]](#).

Note that the total bias is made up of the bias associated with each individual unit i . Thus, each treated observation contributes $\mu^0(X_i) - \mu^0(X_{j(i)})$ to the overall bias. The bias-corrected matching is the following estimator:

$$\hat{\delta}_{ATT}^{BC} = \frac{1}{N_T} \sum_{D_i=1} \left[(Y_i - Y_{j(i)}) - (\hat{\mu}^0(X_i) - \hat{\mu}^0(X_{j(i)})) \right]$$

where $\hat{\mu}^0(X)$ is an estimate of $E[Y|X = x, D = 0]$ using, for example, OLS. Again, I find it always helpful if we take a crack at these estimators with concrete data. Table 21 contains more make-believe data for 8 units, 4 of whom are treated and the rest of whom are functioning as controls. According to the switching equation, we only observe the actual outcomes associated with the potential outcomes under treatment or control, which therefore means we're missing the control values for our treatment group.

Unit	Y^1	Y^0	D	X
1	5		1	11
2	2		1	7
3	10		1	5
4	6		1	3
5		4	0	10
6		0	0	8
7		5	0	4
8		1	0	1

Table 21: Another matching example (this time to illustrate bias correction)

Notice in this example, we cannot implement exact matching because none of the treatment group units have exact matches in the control group. It's worth emphasizing that this is a consequence of finite samples; the likelihood of finding an exact match grows when the sample size of the control group grows faster than that of the treatment group. Instead, we use nearest neighbor matching, which is simply going to be the matching, to each treatment unit, the control group unit whose covariate value is *nearest* to that of the

treatment group unit itself. But, when we do this kind of matching, we necessarily create *matching discrepancies*, which is simply another way of saying that the covariates are not perfectly matched for every unit. Nonetheless, the nearest neighbor “algorithm” creates Table 22.

Unit	Y^1	Y^0	D	X
1	5	4	1	11
2	2	0	1	7
3	10	5	1	5
4	6	1	1	3
5		4	0	10
6		0	0	8
7		5	0	4
8		1	0	1

Table 22: Nearest neighbor matched sample

Recall that the $\widehat{\delta}_{ATT} = \frac{5-4}{4} + \frac{2-0}{4} + \frac{10-5}{4} + \frac{6-1}{4} = 3.25$. With the bias correction, we need to estimate $\widehat{\mu}^0(X)$.⁷⁴ We'll use OLS. Let's illustrate this using another Stata dataset based on Table 22.

⁷⁴ Hopefully, now it will be obvious what exactly $\widehat{\mu}^0(X)$ is. All that it is is the fitted values from a regression of Y on X .

```
. scuse training_bias_reduction, clear
. reg Y X
. predict muhat
. list
```

When we regress Y onto X and D , we get the following estimated coefficients:

$$\begin{aligned}\widehat{\mu}^0(X) &= \widehat{\beta}_0 + \widehat{\beta}_1 X \\ &= 4.42 - 0.49X\end{aligned}$$

This give us the following table of outcomes, treatment status and predicted values.

And then this would be done for the other three simple differences, each of which is added to a bias correction term based on the fitted values from the covariate values.

Now care must be given when using the fitted values for bias correction, so let me walk you through it. You are still going to be taking the simple differences (e.g., 5-4 for row 1), but now you will also subtract out the fitted values associated with each observation's unique covariate. So for instance, in row 1, the outcome 5 has a covariate of 11, which gives it a fitted value of 3.89, but the counterfactual has a value of 10 which gives it a predicted value of 3.94. So therefore we

Unit	Y^1	Y^0	Y	D	X	$\hat{\mu}^0(X)$
1	5	4	5	1	11	3.89
2	2	0	2	1	7	4.08
3	10	5	10	1	5	4.18
4	6	1	6	1	3	4.28
5		4	4	0	10	3.94
6		0	0	0	8	4.03
7		5	5	0	4	4.23
8		1	0	1		4.37

Table 23: Nearest neighbor matched sample with fitted values for bias correction

would use the following bias correction:

$$\widehat{\delta}_{ATT}^{BC} = \frac{5 - 4 - (3.89 - 3.94)}{4} + \dots$$

Now that we see how a specific fitted value is calculated and how it contributes to the calculation of the ATT, let's look at the entire calculation now.

$$\begin{aligned} \widehat{\delta}_{ATT}^{BC} &= \frac{(5 - 4) - (\widehat{\mu}^0(11) - \widehat{\mu}^0(10))}{4} + \frac{(2 - 0) - (\widehat{\mu}^0(7) - \widehat{\mu}^0(8))}{4} \\ &\quad + \frac{(10 - 5) - (\widehat{\mu}^0(5) - \widehat{\mu}^0(4))}{4} + \frac{(6 - 1) - (\widehat{\mu}^0(3) - \widehat{\mu}^0(1))}{4} \\ &= 3.28 \end{aligned}$$

which is slightly higher than the unadjusted ATE of 3.25. Note that this bias correction adjustment becomes more significant as the matching discrepancies themselves become more common. But, if the matching discrepancies are not very common in the first place, then practically by definition, then bias adjustment doesn't change the estimated parameter very much.

Bias arises because of the effect of large matching discrepancies. To minimize these discrepancies, we need a small number of M (e.g., $M = 1$). Larger values of M produce large matching discrepancies. Second, we need matching with replacement. Because matching with replacement can use untreated units as a match more than once, matching with replacement produces smaller discrepancies. And finally, try to match covariates with a large effect on $\mu^0(\cdot)$ well.

The matching estimators have a normal distribution in large samples provided the bias is small. For matching without replacement, the usual variance estimator is valid. That is:

$$\widehat{\sigma}_{ATT}^2 = \frac{1}{N_T} \sum_{D_i=1} \left(Y_i - \frac{1}{M} \sum_{m=1}^M Y_{j_m(i)} - \widehat{\delta}_{ATT} \right)^2$$

For matching with replacement:

$$\begin{aligned}\hat{\sigma}_{ATT}^2 &= \frac{1}{N_T} \sum_{D_i=1} \left(Y_i - \frac{1}{M} \sum_{m=1}^M Y_{j_m(i)} - \hat{\delta}_{ATT} \right)^2 \\ &\quad + \frac{1}{N_T} \sum_{D_i=0} \left(\frac{K_i(K_i - 1)}{M^2} \right) \widehat{var}(\varepsilon | X_i, D_i = 0)\end{aligned}$$

where K_i is the number of times that observation i is used as a match. $\widehat{var}(Y_i | X_i, D_i = 0)$ can be estimated also by matching. For example, take two observations with $D_i = D_j = 0$ and $X_i \approx X_j$, then

$$\widehat{var}(Y_i | X_i, D_i = 0) = \frac{(Y_i - Y_j)^2}{2}$$

is an unbiased estimator of $\widehat{var}(\varepsilon_i | X_i, D_i = 0)$. The bootstrap, though, doesn't work.

Propensity score methods There are several ways of achieving the conditioning strategy implied by the backdoor criterion. One additional one was developed by Donald Rubin in the mid-1970s to early 1980s called the propensity score method [Rubin, 1977, Rosenbaum and Rubin, 1983]. The propensity score is very similar in spirit to both nearest neighbor covariate matching by Abadie and Imbens [2006] and subclassification. It's a very popular method, particularly in the medical sciences, of addressing selection on observables and has gained some use among economists as well [Dehejia and Wahba, 2002].

Before we dig into it, though, a couple of words to help manage your expectations. Propensity score matching has not been as widely used by economists as other methods for causal inference because economists are oftentimes skeptical that CIA can be achieved in any dataset. This is because for many applications, economists as a group are more concerned about selection on unobservables than they are of selection on observables, and as such, have not found matching methods to be used as often. I am agnostic as to whether CIA holds or doesn't in your particular application, though. Only a DAG will tell you what the appropriate identification strategy is, and insofar as the backdoor criterion can be met, then matching methods may be appropriate.

Propensity score matching is used when treatment is nonrandom but is believed to be based on a variety of observable covariates. It requires that the CIA hold in the data. Propensity score matching takes those covariates needed to satisfy CIA, estimates a maximum likelihood model of the conditional probability of treatment, and uses the predicted values from that estimation to collapse those covariates

into a single scalar. All comparisons between the treatment and control group are then based on that value.⁷⁵ But I cannot emphasize this enough – this method, like regression more generally, only has value for your project if you can satisfy the backdoor criterion by conditioning on X . If you cannot satisfy the backdoor criterion in your data, then the propensity score gains you nothing. It is absolutely critical that your DAG be, in other words, defensible and accurate, as you depend on those theoretical relationships to design the appropriate identification strategy.⁷⁶

The idea with propensity score methods is to compare units who, based on observables, had very similar probabilities of being placed into the treatment group even though those units differed with regards to actual treatment assignment. If conditional on X , two units have the same probability of being treated, then we say they have similar *propensity scores*. If two units have the same propensity score, but one is the treatment group and the other is not, and the *conditional independence assumption* (CIA) credibly holds in the data, then differences between their observed outcomes are attributable to the treatment. CIA in this context means that the assignment of treatment, conditional on the propensity score, is independent of potential outcomes, or “as good as random”.⁷⁷

One of the goals when using propensity score methods is to create covariate balance between the treatment group and control group such that the two groups become observationally *exchangeable*.⁷⁸ There are three steps to using propensity score matching. The first step is to estimate the propensity score; the second step is to select an algorithmic method incorporating the propensity score to calculate average treatment effects; the final step is to calculate standard errors. The first is always the same regardless of which algorithmic method we use in the second stage: we use maximum likelihood models to estimate the conditional probability of treatment, usually probit or logit. Before walking through an example using real data, let’s review some papers that use it.

Example: the NSW Job Trainings Program The National Supported Work Demonstration (NSW) job trainings program was operated by the Manpower Demonstration Research Corp (MRDC) in the mid-1970s. The NSW was a temporary employment program designed to help disadvantaged workers lacking basic job skills move into the labor market by giving them work experience and counseling in a sheltered environment. It was also unique in that it **randomly assigned** qualified applicants to training positions. The treatment group received all the benefits of the NSW program. The controls were basically left to fend for themselves. The program admitted

⁷⁵ There are multiple methods that use the propensity score, as we will see, but they all involve using the propensity score to make valid comparisons between the treatment group and control group.

⁷⁶ We will discuss in the instrumental variables chapter a common method for addressing a situation where the backdoor criterion cannot be met in your data.

⁷⁷ This is what meant by the phrase *selection on observables*.

⁷⁸ Exchangeable simply means that the two groups *appear* similar to one another on *observables*.

AFDC females, ex-drug addicts, ex-criminal offenders, and high school dropouts of both sexes.

Treatment group members were guaranteed a job for 9-18 months depending on the target group and site. They were then divided into crews of 3-5 participants who worked together and met frequently with an NSW counselor to discuss grievances and performance. Finally, they were paid for their work. NSW offered the trainees lower wage rates than they would've received on a regular job, but allowed their earnings to increase for satisfactory performance and attendance. After their term expired, they were forced to find regular employment. The kinds of jobs varied within sites – some were gas station attendants, some worked at a printer shop – and males and females were frequently performing different kinds of work.

The MDRC collected earnings and demographic information from both the treatment and the control group at baseline as well as every 9 months thereafter. MDRC also conducted up to 4 post-baseline interviews. There were different sample sizes from study to study, which can be confusing, but it has simple explanations.

NSW was a randomized job trainings program; therefore the independence assumption was satisfied. So calculating average treatment effects was straightforward – it's the simple difference in means estimator that we discussed in the Rubin causal chapter.⁷⁹

$$\frac{1}{N_T} \sum_{D_i=1} Y_i - \frac{1}{N_C} \sum_{D_i=0} Y_i \approx E[Y^1 - Y^0]$$

The good news for MDRC, and the treatment group, was that the treatment worked.⁸⁰ Treatment group participants' real earnings post-treatment in 1978 was larger than that of the control group by approximately \$900 [Lalonde, 1986] to \$1,800 [Dehejia and Wahba, 2002], depending on the sample the researcher used.

Lalonde [1986] is an interesting study both because he is evaluating the NSW program, and because he is evaluating commonly used econometric methods from that time. He evaluated the econometric estimators' performance by trading out the experimental control group data with non-experimental control group data drawn from the population of US citizens. He used three samples of the Current Population Survey (CPS) and three samples of the Panel Survey of Income Dynamics (PSID) for this non-experimental control group data. Non-experimental data is, after all, the typical situation an economist finds herself in. But the difference with the NSW is that it was a randomized experiment, and therefore we know the average treatment effect. Since we know the average treatment effect, we can see how well a variety of econometric models perform. If the NSW program increased earnings by $\approx \$900$, then we should find that if

⁷⁹ Remember, randomization means that the treatment was independent of the potential outcomes, so simple difference in means identifies the average treatment effect.

⁸⁰ Lalonde [1986] lists several studies that discuss the findings from the program in footnote 3.

the other econometrics estimators does a good job, right?

Lalonde [1986] reviewed a number of popular econometric methods from this time using both the PSID and the CPS samples as non-experimental comparison groups, and his results were consistently bad. Not only were his estimates usually very different in magnitude, but his results are almost always the wrong sign! This paper, and its pessimistic conclusion, was influential in policy circles and led to greater push for more experimental evaluations.⁸¹ We can see these results in the following tables from Lalonde [1986]. Figure 17 shows the effect of the treatment when comparing the treatment group to the experimental control group. The baseline difference in real earnings between the two groups were negligible,⁸² But the post-treatment difference in average earnings was between \$798 and \$886.

TABLE 5—EARNINGS COMPARISONS AND ESTIMATED TRAINING EFFECTS FOR THE NSW MALE PARTICIPANTS USING COMPARISON GROUPS FROM THE PSID AND THE CPS-SSA^{a,b}

Name of Comparison Group ^d	Comparison Group Earnings Growth 1975–78 (1)	NSW Treatment Earnings Less Comparison Group Earnings				Difference in Differences: Difference in Earnings Growth 1975–78 Treatments Less Comparisons		Unrestricted Difference in Differences: Quasi Difference in Earnings Growth 1975–78		Controlling for All Observed Variables and Pre-Training Earnings (10)
		Pre-Training Year, 1975		Post-Training Year, 1978		Without Age (6)	With Age (7)	Unadjusted (8)	Adjusted ^c (9)	
		Unadjusted ^c (2)	Adjusted ^c (3)	Unadjusted ^c (4)	Adjusted ^c (5)					
Controls	\$2,063 (325)	\$39 (383)	\$–21 (378)	\$886 (476)	\$798 (472)	\$847 (560)	\$856 (558)	\$897 (467)	\$802 (467)	\$662 (506)
PSID-1	\$2,043 (237)	–\$15,997 (795)	–\$7,624 (851)	–\$15,578 (913)	–\$8,067 (990)	\$425 (650)	–\$749 (692)	–\$2,380 (680)	–\$2,119 (746)	–\$1,228 (896)
PSID-2	\$6,071 (637)	–\$4,503 (608)	–\$3,669 (757)	–\$4,020 (781)	–\$3,482 (935)	\$484 (738)	–\$650 (850)	–\$1,364 (729)	–\$1,694 (878)	–\$792 (1024)
PSID-3	(\$3,322 (780))	(\$455 (539))	(\$455 (704))	(\$697 (760))	(\$509 (967))	\$242 (884)	–\$1,325 (1078)	\$629 (757)	–\$552 (967)	\$397 (1103)
CPS-SSA-1	\$1,196 (61)	–\$10,585 (539)	–\$4,654 (509)	–\$8,870 (562)	–\$4,416 (557)	\$1,714 (452)	\$195 (441)	–\$1,543 (426)	–\$1,102 (450)	–\$805 (484)
CPS-SSA-2	\$2,684 (229)	–\$4,321 (450)	–\$1,824 (535)	–\$4,095 (537)	–\$1,675 (672)	\$226 (539)	–\$488 (530)	–\$1,850 (497)	–\$782 (621)	–\$319 (761)
CPS-SSA-3	\$4,548 (409)	\$337 (343)	\$878 (447)	–\$1,300 (590)	\$224 (766)	–\$1,637 (631)	–\$1,388 (655)	–\$1,396 (582)	\$17 (761)	\$1,466 (984)

^aThe columns above present the estimated training effect for each econometric model and comparison group. The dependent variable is earnings in 1978. Based on the experimental data an unbiased estimate of the impact of training presented in col. 4 is \$886. The first three columns present the difference between each comparison group's 1975 and 1978 earnings and the difference between the pre-training earnings of each comparison group and the NSW treatments.

^bEstimates are in 1982 dollars. The numbers in parentheses are the standard errors.

^cThe exogenous variables used in the regression adjusted equations are age, age squared, years of schooling, high school dropout status, and race.

^dSee Table 3 for definitions of the comparison groups.

Figure 18 shows the results he got when he used the non-experimental data as the comparison group. He used three samples of the PSID and three samples of the CPS. In nearly every point estimate, the effect is negative.

So why is there such a stark difference when we move from the NSW control group to either the PSID or CPS? The reason is because of selection bias. That is

$$E[Y^0 | D = 1] \neq E[Y^0 | D = 0]$$

In other words, it's highly likely that the real earnings of NSW par-

⁸¹ It's since been cited a little more than 1,700 times.

⁸² The treatment group made \$39 more than the control group in the simple difference and \$21 less in the multivariate regression model, but neither is statistically significant.

Figure 17: Lalonde [1986] Table 5(a)

TABLE 5—EARNINGS COMPARISONS AND ESTIMATED TRAINING EFFECTS FOR THE NSW
MALE PARTICIPANTS USING COMPARISON GROUPS FROM THE PSID AND THE CPS-SSA^{a,b}

Name of Comparison Group ^d	Comparison Group Earnings Growth 1975–78 (1)	NSW Treatment Earnings Less Comparison Group Earnings				Difference in Differences: Difference in Earnings		Unrestricted Difference in Differences: Quasi Difference in Earnings Growth 1975–78		Controlling for All Observed Variables and Pre-Training Earnings (10)	
		Pre-Training Year, 1975		Post-Training Year, 1978		Growth 1975–78 Treatments Less Comparisons		Without Age (6) With Age (7)			
		Unadjusted ^e (2)	Adjusted ^c (3)	Unadjusted ^e (4)	Adjusted ^c (5)	Without Age (6)	With Age (7)	Unadjusted ^e (8)	Adjusted ^c (9)		
Controls	\$2,063 (325)	\$39 (383)	-\$21 (378)	\$886 (476)	\$798 (472)	\$847 (560)	\$856 (558)	\$897 (467)	\$802 (467)	\$662 (506)	
PSID-1	\$2,043 (237)	-\$15,997 (795)	-\$7,624 (851)	-\$15,578 (913)	-\$8,067 (990)	\$425 (650)	-\$749 (692)	-\$2,380 (680)	-\$2,119 (746)	-\$1,228 (896)	
PSID-2	\$6,071 (637)	-\$4,503 (608)	-\$3,669 (757)	-\$4,020 (781)	-\$3,482 (935)	\$484 (738)	-\$650 (850)	-\$1,364 (729)	-\$1,694 (878)	-\$792 (1024)	
PSID-3	\$3,322 (780)	(\$455 (539))	\$455 (704)	\$697 (760)	-\$509 (967)	\$242 (884)	-\$1,325 (1078)	\$629 (757)	-\$552 (967)	\$397 (1103)	
CPS-SSA-1	\$1,196 (61)	-\$10,585 (539)	-\$4,654 (509)	-\$8,870 (562)	-\$4,416 (557)	\$1,714 (452)	\$195 (441)	-\$1,543 (426)	-\$1,102 (450)	-\$805 (484)	
CPS-SSA-2	\$2,684 (229)	-\$4,321 (450)	-\$1,824 (535)	-\$4,095 (537)	-\$1,675 (672)	\$226 (539)	-\$488 (530)	-\$1,850 (497)	-\$782 (621)	-\$319 (761)	
CPS-SSA-3	\$4,548 (409)	\$337 (343)	\$878 (447)	-\$1,300 (590)	\$224 (766)	-\$1,637 (631)	-\$1,388 (655)	-\$1,396 (582)	\$17 (761)	\$1,466 (984)	

^aThe columns above present the estimated training effect for each econometric model and comparison group. The dependent variable is earnings in 1978. Based on the experimental data an unbiased estimate of the impact of training presented in col. 4 is \$886. The first three columns present the difference between each comparison group's 1975 and 1978 earnings and the difference between the pre-training earnings of each comparison group and the NSW treatments.

^bEstimates are in 1982 dollars. The numbers in parentheses are the standard errors.

^cThe exogenous variables used in the regression adjusted equations are age, age squared, years of schooling, high school dropout status, and race.

^dSee Table 3 for definitions of the comparison groups.

ticipants would have been much lower than the non-experimental control group's earnings. As you recall from our decomposition of the simple difference in means estimator, the second form of bias is selection bias, and if $E[Y^0|D = 1] < E[Y^0|D = 0]$, this will bias the estimate of the ATE downward (e.g., estimates that show a negative effect).

But a violation of independence also implies that the balancing property doesn't hold. Table 24 shows the mean values for each covariate for the treatment and control groups where the control is the 15,992 observations from the CPS. As you can see, the treatment group appears to be very different on average from the control group CPS sample along nearly every covariate listed. The NSW participants are more black, more hispanic, younger, less likely to be married, more likely to have no degree, less schooling, more likely to be unemployed in 1975 and have considerably lower earnings in 1975. In short, the two groups are not *exchangeable* on observables (and likely not exchangeable on unobservables either).

The first paper to re-evaluate Lalonde [1986] using propensity score methods was Dehejia and Wahba [1999].⁸³ Their interest was two fold - to examine whether propensity score matching could be an improvement in estimating treatment effects using non-experimental data. And two, to show the diagnostic value of propensity score matching. The authors used the same non-experimental control group datasets from the CPS and PSID as Lalonde [1986].

Figure 18: Lalonde [1986] Table 5(b)

⁸³ Lalonde [1986] did not review propensity score matching in this study. One possibility is that he wasn't too familiar with the method. Rosenbaum and Rubin [1983] was relatively new, after all, when LaLonde had begun his project and had not yet been incorporated into most economists' toolkit.

Table 24: Completed matching example with single covariate

covariate	All		CPS	NSW		
	mean	(s.d.)	Controls	Trainees	t-stat	diff
			$N_c = 15,992$	$N_t = 297$		
Black	0.09	0.28	0.07	0.80	47.04	-0.73
Hispanic	0.07	0.26	0.07	0.94	1.47	-0.02
Age	33.07	11.04	33.2	24.63	13.37	8.6
Married	0.70	0.46	0.71	0.17	20.54	0.54
No degree	0.30	0.46	0.30	0.73	16.27	-0.43
Education	12.0	2.86	12.03	10.38	9.85	1.65
1975 Earnings	13.51	9.31	13.65	3.1	19.63	10.6
1975 Unemp	0.11	0.32	0.11	0.37	14.29	-0.26

Let's walk through what the authors did in steps, and what they learned from each of these steps. First, the authors estimated the propensity score. Then the authors sought to create balance on observable covariates through trimming. By trimming I mean that the authors discarded control units with propensity score values outside the range of the treatment group in order to impose common support.

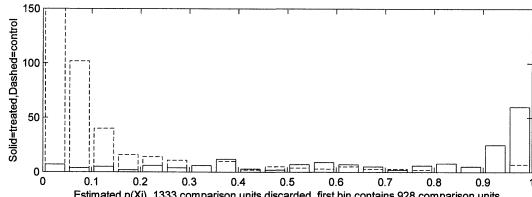


Figure 1. Histogram of the Estimated Propensity Score for NSW Treated Units and PSID Comparison Units. The 1,333 PSID units whose estimated propensity score is less than the minimum estimated propensity score for the treatment group are discarded. The first bin contains 928 PSID units. There is minimal overlap between the two groups. Three bins (.8-.85, .85-.9, and .9-.95) contain no comparison units. There are 97 treated units with an estimated propensity score greater than .8 and only 7 comparison units.

Figure 19: Dehejia and Wahba [1999]
Figure 1, overlap in the propensity scores (using PSID)

Fig 19 shows the overlap in the propensity score for the treatment and control group units using the PSID comparison units. 1,333 comparison units were discarded because they fell outside the range of the treatment group's propensity score distribution. As can be seen, there is a significant number of PSID control group units with very low propensity score values – 928 comparison units are contained in the first bin. This is not surprising because, after all, most of the population differed significantly from the treatment group, and thus had a low probability of treatment.

Next the authors do the same for the CPS sample. Here the overlap is even worse. They dropped 12,611 observations in the control

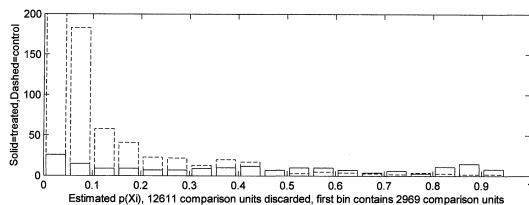


Figure 2. Histogram of the Estimated Propensity Score for NSW Treated Units and CPS Comparison Units. The 12,611 CPS units whose estimated propensity score is less than the minimum estimated propensity score for the treatment group are discarded. The first bin contains 2,969 CPS units. There is minimal overlap between the two groups, but the overlap is greater than in Figure 1; only one bin (.45-.5) contains no comparison units, and there are 35 treated and 7 comparison units with an estimated propensity score greater than .8.

group because their propensity scores were outside the treatment group range. Also, a large number of observations have low propensity scores, evidenced by the fact that the first bin contains 2,969 comparison units. While there is minimal overlap between the two groups, the overlap is greater in Figure 20 than Figure 19.

Table 3. Estimated Training Effects for the NSW Male Participants Using Comparison Groups From PSID and CPS

	NSW earnings less comparison group earnings		NSW treatment earnings less comparison group earnings, conditional on the estimated propensity score					
			Quadratic in score ^b	Stratifying on the score			Matching on the score	
	(1) Unadjusted	(2) Adjusted ^a		(4) Unadjusted	(5) Adjusted	(6) Observations ^c	(7) Unadjusted	(8) Adjusted ^d
NSW	1,794 (633)	1,672 (638)						
PSID-1 ^e	-15,205 (1,154)	731 (886)	294 (1,389)	1,608 (1,571)	1,494 (1,581)	1,255 (2,235)	1,691 (2,209)	1,473 (809)
PSID-2 ^f	-3,647 (959)	683 (1,028)	496 (1,193)	2,220 (1,768)	2,321 (1,793)	389 (2,335)	1,455 (2,303)	1,480 (808)
PSID-3 ^g	1,069 (899)	825 (1,104)	647 (1,383)	1,870 (1,994)	1,870 (2,002)	247 (2,335)	2,120 (826)	1,549 (826)
CPS-1 ^h	-8,498 (712)	972 (550)	1,117 (747)	1,713 (1,115)	1,774 (1,152)	4,117 (2,335)	1,582 (1,069)	1,616 (751)
CPS-2 ^h	-3,822 (670)	790 (658)	505 (847)	1,543 (1,461)	1,622 (1,346)	1,493 (2,025)	1,788 (1,205)	1,563 (753)
CPS-3 ^h	-635 (657)	1,326 (798)	556 (951)	1,252 (1,617)	2,219 (2,082)	514 (1,496)	587 (776)	662 (776)

^a Least squares regression: RET78 on a constant, a treatment indicator, age, age², education, no degree, black, Hispanic, RET4, RET75.

^b Least squares regression of RET78 on a quadratic on the estimated propensity score and a treatment indicator, for observations used under stratification; see note (g).

^c Number of observations refers to the actual number of comparison and treatment units used for (3)-(5); namely, all treatment units and those comparison units whose estimated propensity score is greater than the minimum, and less than the maximum, estimated propensity score for the treatment group.

^d Weighted least squares: treatment observations weighted as 1, and control observations weighted by the number of times they are matched to a treatment observation [same covariates as (a)].

Propensity scores are estimated using the logistic model, with specifications as follows:

^e PSID-1: Prob ($T_i = 1$) = $F(\text{age}, \text{age}^2, \text{education}, \text{education}^2, \text{married}, \text{no degree}, \text{black}, \text{Hispanic}, \text{RET4}, \text{RET5}, \text{RET4}^2, \text{RET5}^2, u74/\text{black})$.

^f PSID-2 and PSID-3: Prob ($T_i = 1$) = $F(\text{age}, \text{age}^2, \text{education}, \text{education}^2, \text{no degree}, \text{married}, \text{black}, \text{Hispanic}, \text{RET4}, \text{RET5}^2, \text{RET5}^3, u74, u75)$.

^g CPS-1, CPS-2, and CPS-3: Prob ($T_i = 1$) = $F(\text{age}, \text{age}^2, \text{education}, \text{education}^2, \text{no degree}, \text{married}, \text{black}, \text{Hispanic}, \text{RET4}, \text{RET5}, \text{u74}, \text{u75}, \text{education}^*\text{RET4}, \text{age}^3)$.

Figure 21 shows the results using propensity score weighting/matching.⁸⁴ As can be seen, the results are a considerable improvement over Lalonde [1986]. I won't review every treatment effect they estimate, but I will note that they are all positive and similar in magnitude to what they found in columns 1-2 using only the experimental data.

Finally, the authors examined the balance between the covariates in the treatment group (NSW) and the various non-experimental (matched) samples. Recall that the balancing property suggests that covariate values will be the same for treatment and control group after they trim the outlier propensity score units from the data. Figure 22 shows the sample means of characteristics in the matched control sample versus the experimental NSW sample (first row).

Figure 20: Dehejia and Wahba [1999]
Figure 2, overlap in the propensity scores (using CPS)

Figure 21: Dehejia and Wahba [1999]
Table 3 results.

Let's hold off digging into exactly how they used the propensity score to generate these estimates.

Table 4. Sample Means of Characteristics for Matched Control Samples

Matched samples	No. of observations	Age	Education	Black	Hispanic	No degree	Married	RE74 (U.S. \$)	RE75 (U.S. \$)
NSW	185	25.81	10.35	.84	.06	.71	.19	2,096	1,532
MPSID-1	56	26.39	10.62	.86	.02	.55	.15	1,794	1,126
		[2.56]	[.63]	[.13]	[.06]	[.13]	[.12]	[1,406]	[1,146]
MPSID-2	49	25.32	11.10	.89	.02	.57	.19	1,599	2,225
		[2.63]	[.82]	[.14]	[.08]	[.16]	[.16]	[1,905]	[1,228]
MPSID-3	30	26.86	10.96	.91	.01	.52	.25	1,386	1,863
		[2.97]	[.84]	[.13]	[.08]	[.16]	[.16]	[1,680]	[1,494]
MCPS-1	119	26.91	10.52	.86	.04	.64	.19	2,110	1,396
		[1.25]	[.32]	[.06]	[.04]	[.07]	[.06]	[841]	[563]
MCPS-2	87	26.21	10.21	.85	.04	.68	.20	1,758	1,204
		[1.43]	[.37]	[.08]	[.05]	[.09]	[.08]	[896]	[661]
MCPS-3	63	25.94	10.69	.87	.06	.53	.13	2,709	1,587
		[1.68]	[.48]	[.09]	[.06]	[.10]	[.09]	[1,285]	[760]

NOTE: Standard error on the difference in means with NSW sample is given in brackets.
MPSID1-3 and MCPS1-3 are the subsamples of PSID1-3 and CPS1-3 that are matched to the treatment group.

Figure 22: Dehejia and Wahba [1999]
Table 4, covariate balance

Trimming on the propensity score, in effect, satisfied the balancing property. Covariates are much closer in mean value to the NSW sample after trimming on the propensity score.

Estimation Propensity score is best explained, in my opinion as with other methods, using actual data. We will use data from Dehejia and Wahba [2002] for the following exercises. I encourage you to copy these commands into a do file and run them on your own so that you can see the analysis directly. First we need to download the data. Then we will calculate the ATE using the experimental treatment and control units.

```
. scuse nsw_dw
. su re78 if treat
. gen y1 = r(mean)
. su re78 if treat==0
. gen y0 = r(mean)
. gen ate = y1-y0
. su ate
. di 6349.144 - 4554.801
```

Which yields an ATE of \$1,794.343. Next we do the same for the CPS data. We will not include the analysis of the PSID control group for brevity. So first we append the main dataset with the CPS files:

```
. append using "http://scunning.com/teaching/cps_controls"
```

Next we construct the controls discussed in the footnote of Table 2 in Dehejia and Wahba [2002]:

```
. gen agesq=age*age
. gen agecube=age^3
. ren education school
. gen schoolsq=school^2
. gen u74 = 0 if re74!=.
. replace u74 = 1 if re74==0
. gen u75 = 0 if re75!=.
. replace u75 = 1 if re75==0
```

```
. gen interaction1 = school*re74
. gen re74sq=re74^2
. gen re75sq=re75^2
. gen interaction2 = u74*hispanic
```

Now we are ready to estimate the propensity score. We will use a logit model to be consistent with [Dehejia and Wahba \[2002\]](#).

```
. logit treat age agesq agecube school schoolsq married
nodegree black hispanic re74 re75 u74 u75 interaction1
. predict pscore
```

The predict command uses the estimated coefficients from our logit model and then estimates the conditional probability of treatment using:

$$Pr(D = 1|X) = F(\beta_0 + \gamma Treat + \alpha X)$$

where $F() = \frac{e}{(1+e)}$ is the cumulative logistic distribution.

The propensity score used the fitted values from the maximum likelihood regression to calculate each unit's conditional probability of treatment *regardless of their actual treatment status*. The propensity score is just the predicted conditional probability of treatment or fitted value for each unit.⁸⁵

The definition of the propensity score is the selection probability conditional on the confounding variables; $p(X) = Pr(D = 1|X)$. There are two identifying assumptions for propensity score methods. The first is CIA. That is, $(Y^0, Y^1) \perp\!\!\!\perp D|X$. The second is called the *common support assumption*. That is, $0 < Pr(D = 1|X) < 1$ which is the common support assumption.⁸⁶ The conditional independence assumption simply means that the backdoor criterion is met in the data by conditioning on a vector X . Or, put another way, conditional on X , the assignment of units to the treatment is *as good as random*.⁸⁷ This is

$$\begin{aligned} Y_i^0 &= \alpha + \beta X_i + \varepsilon_i \\ Y_i^1 &= Y_i^0 + \delta \\ Y_i &= \alpha + \delta D_i + \beta X_i + \varepsilon_i \end{aligned}$$

Conditional independence is the same as assuming $\varepsilon_i \perp\!\!\!\perp D_i|X_i$. One last thing before we move on: CIA is **not testable** because it requires potential outcomes, which we do not have. We only have observed outcomes according to the switching equation. CIA is an assumption, and it may or may not be a credible assumption depending on your application.

The second identifying assumption is called the *common support assumption*. It is required to calculate any particular kind of defined average treatment effect, and without it, you will just get some kind of weird weighted average treatment effect for only those regions that

⁸⁵ It is advisable to use maximum likelihood when estimating the propensity score because so that the fitted values are in the range $[0, 1]$. We could use a linear probability model, but linear probability models routinely create fitted values below 0 and above 1, which are not true probabilities since $0 \leq p \leq 1$.

⁸⁶ This simply means that for any probability, there must be units in both the treatment group *and* the control group.

⁸⁷ CIA is expressed in different ways according to the econometric/statistical tradition. [Rosenbaum and Rubin \[1983\]](#) called it the ignorable treatment assignment, or *unconfoundedness*. Pearl calls it the backdoor criterion. [Barnow et al. \[1981\]](#) and [Dale and Krueger \[2002\]](#) call it *selection on observables*. In the traditional econometric pedagogy, as we discussed earlier, it's called the zero conditional mean assumption as we see below.

do have common support. Common support requires that for each value of X , there is a positive probability of being both treated and untreated, or $0 < Pr(D_i = 1|X_i) < 1$. This implies that the probability of receiving treatment for every value of the vector X is strictly within the unit interval. Common support ensures there is sufficient overlap in the characteristics of treated and untreated units to find adequate matches. Unlike CIA, the common support requirement is **testable** by simply plotting histograms or summarizing the data. Here we do that two ways: by looking at the summary statistics and looking at a histogram.

```
. su pscore if treat==1, detail
```

Treatment group		
Percentiles	Values	Smallest
1%	.0022114	.0018834
5%	.0072913	.0022114
10%	.0202463	.0034608
25%	.0974065	.0035149
50%	.1938186	
Percentiles	Values	Largest
75%	.3106517	.5583002
90%	.4760989	.5698137
95%	.5134488	.5705917
99%	.5705917	.5727966

Table 25: Distribution of propensity score for treatment group.

```
. su pscore if treat==0, detail
```

CPS Control group		
Percentiles	Values	Smallest
1%	5.14e-06	4.69e-07
5%	.0000116	6.79e-07
10%	.0000205	7.78e-07
25%	.0000681	9.12e-07
50%	.0003544	
Percentiles	Values	Largest
75%	.0021622	.5727351
90%	.0085296	.5794474
95%	.0263618	.5929902
99%	.2400503	.5947019

Table 26: Distribution of propensity score for CPS Control group.

The mean value of the propensity score for the treatment group is 0.22 and the mean for the CPS control group is 0.009. The 50th

percentile for the treatment group is 0.194 but the control group doesn't reach that high a number until almost the 99th percentile. Let's look at the distribution of the propensity score for the two groups using a histogram now.

```
. histogram pscore, by(treat) binrescale
```

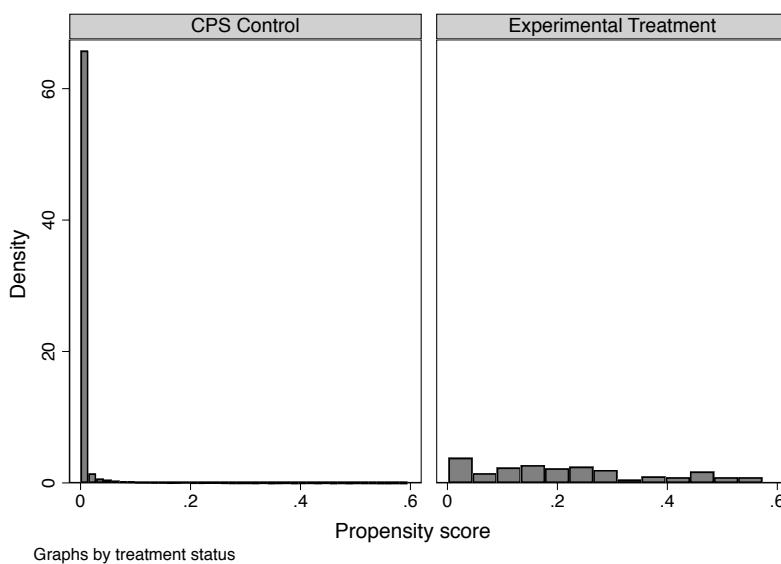


Figure 23: Histogram of propensity score by treatment status

These two simple diagnostic tests show what is going to be a problem later when we use inverse probability weighting. The probability of treatment is spread out across the units in the treatment group, but there is a very large mass of nearly zero propensity scores in the CPS. How do we interpret this? What this means is that the characteristics of individuals in the treatment group are rare in the CPS sample. This is not surprising given the strong negative selection into treatment. These individuals are younger, less likely to be married, and more likely to be uneducated and a minority. The lesson is if the two groups are significantly different on background characteristics, then the propensity scores will have grossly different distributions by treatment status. We will discuss this in greater detail later.

For now, let's look at the treatment parameter under both assumptions.

$$\begin{aligned} E[\delta_i(X_i)] &= E[Y_i^1 - Y_i^0 | X_i = x] \\ &= E[Y_i^1 | X_i = x] - E[Y_i^0 | X_i = x] \end{aligned}$$

The conditional independence assumption allows us to make the

following substitution

$$E[Y_i^1 | D_i = 1, X_i = x] = E[Y_i | D_i = 1, X_i = x]$$

and same for the other term. Common support means we can estimate both terms. Therefore under both assumptions

$$\delta = E[\delta(X_i)]$$

From these assumptions we get the *propensity score theorem*, which states that if $(Y^1, Y^0) \perp\!\!\!\perp D|X$ (CIA) then $(Y^1, Y^0) \perp\!\!\!\perp D|p(X)$ where $p(X) = Pr(D = 1|X)$, the propensity score. This means that conditioning on the propensity score is sufficient to have independence. Conditioning on the propensity score is enough to have independence between the treatment and the potential outcomes.

This is an extremely valuable theorem because stratifying on X tends to run into the sparseness-related problems (i.e., empty cells) in finite samples for even a moderate number of covariates. But the propensity scores is just a scalar. So stratifying across a probability is going to reduce that dimensionality problem.

The proof of the propensity score theorem is fairly straightforward as it's just an application of the law of iterated expectations with nested conditioning.⁸⁸ If we can show that the probability an individual receives treatment conditional on potential outcomes and the propensity score is not a function of potential outcomes, then we will have proven that there is independence between the potential outcomes and the treatment conditional on X . Before diving into the proof, first recognize that

$$Pr(D = 1|Y^0, Y^1, p(X)) = E[D|Y^0, Y^1, p(X)]$$

because

$$\begin{aligned} E[D|Y^0, Y^1, p(X)] &= 1 \times Pr(D = 1|Y^0, Y^1, p(X)) \\ &\quad + 0 \times Pr(D = 0|Y^0, Y^1, p(X)) \end{aligned}$$

and the second term cancels out because it's multiplied zero. The

⁸⁸ See Angrist and Pischke [2009] p. 80-81.

formal proof is as follows:

$$\begin{aligned}
 Pr(D = 1|Y^1, Y^0, p(X)) &= \underbrace{E[D|Y^1, Y^0, p(X)]}_{\text{See previous description}} \\
 &= \underbrace{E[E[D|Y^1, Y^0, p(X), X]|Y^1, Y^0, p(X)]}_{\text{by LIE}} \\
 &= \underbrace{E[E[D|Y^1, Y^0, X]|Y^1, Y^0, p(X)]}_{\text{Given } X, \text{ we know } p(X)} \\
 &= \underbrace{E[E[D|X]|Y^1, Y^0, p(X)]}_{\text{by conditional independence}} \\
 &= \underbrace{E[p(X)|Y^1, Y^0, p(X)]}_{\text{propensity score definition}} \\
 &= p(X)
 \end{aligned}$$

Using a similar argument, we obtain:

$$\begin{aligned}
 Pr(D = 1|p(X)) &= \underbrace{E[D|p(X)]}_{\text{Previous argument}} \\
 &= \underbrace{E[E[D|X]|p(X)]}_{\text{LIE}} \\
 &= \underbrace{E[p(X)|p(X)]}_{\text{definition}} \\
 &= p(X)
 \end{aligned}$$

and $Pr(D = 1|Y^1, Y^0, p(X)) = Pr(D = 1|p(X))$ by CIA.

Like the omitted variable bias formula for regression, the propensity score theorem says that you need only control for covariates that determine the likelihood a unit receives the treatment. But it also says something more than that. It technically says that the *only* covariate you need to condition on is the propensity score. All of the information from the X matrix has been collapsed into a single number: the propensity score.

A corollary of the propensity score theorem, therefore, states that given CIA, we can estimate average treatment effects by weighting appropriately the simple difference in means.⁸⁹

Because the propensity score is a function of X , we know

$$\begin{aligned}
 Pr(D = 1|X, p(X)) &= Pr(D = 1|X) \\
 &= p(X)
 \end{aligned}$$

Therefore conditional on the propensity score, the probability that $D = 1$ does not depend on X any longer. That is, D and X are independent of one another conditional on the propensity score, or

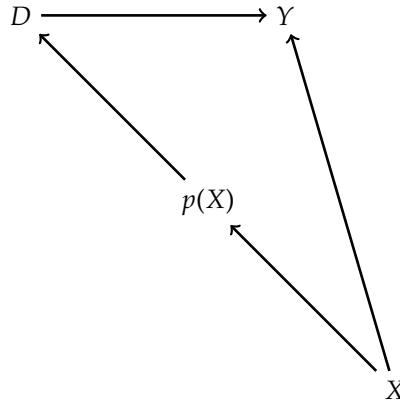
$$D \perp\!\!\!\perp |p(X)$$

⁸⁹ This all works if we match on the propensity score and then calculate differences in means. Direct propensity score matching works in the same way as the covariate matching we discussed earlier (e.g., nearest neighbor matching) except that we match on the *score* instead of the *covariates* directly.

So from this we also obtain the *balancing property* of the propensity score:

$$\Pr(X|D = 1, p(X)) = \Pr(X|D = 0, p(X))$$

which states that conditional on the propensity score, the distribution of the covariates is the same for treatment as it is for control group units. See this in the following DAG.



Notice that there exists two paths between X and D . There's the direct path of $X \rightarrow p(X) \rightarrow D$ and there's the backdoor path $X \rightarrow Y \leftarrow D$. The backdoor path is blocked by a collider, so there is not systematic correlation between X and D through it. But there is systematic correlation between X and D through the first directed path. But, when we condition on $p(X)$, the propensity score, notice that D and X are statistically *independent*. This implies that $D \perp\!\!\!\perp X|p(X)$ which implies

$$\Pr(X|D = 1, \hat{p}(X)) = \Pr(X|D = 0, \hat{p}(X))$$

This is something we can directly test, but note the implication: conditional on the propensity score, treatment and control should on average be the same with respect to X . In other words, the propensity score theorem implies *balanced* observable covariates.⁹⁰

Estimation using propensity score matching

Inverse probability weighting has become a common approach within the context of propensity score estimation. We have the following proposition related to weighting. If CIA holds, then

$$\begin{aligned}\delta_{ATE} &= E[Y^1 - Y^0] \\ &= E\left[Y \cdot \frac{D - p(X)}{p(X) \cdot (1 - p(X))}\right] \\ \delta_{ATT} &= E[Y^1 - Y^0 | D = 1] \\ &= \frac{1}{\Pr(D = 1)} \cdot E\left[Y \cdot \frac{D - p(X)}{1 - p(X)}\right]\end{aligned}$$

⁹⁰ I will have now officially beaten the dead horse. But please understand - just because something is exchangeable on observables does not make it exchangeable on unobservables. The propensity score theorem does *not* imply balanced unobserved covariates. See Brooks and Ohsfeldt [2013].

The proof for this is:

$$\begin{aligned} E \left[Y \frac{D - p(X)}{p(X)(1 - p(X))} \middle| X \right] &= E \left[\frac{Y}{p(X)} \middle| X, D = 1 \right] p(X) \\ &\quad + E \left[\frac{-Y}{1 - p(X)} \middle| X, D = 0 \right] (1 - p(X)) \\ &= E[Y|X, D = 1] - E[Y|X, D = 0] \end{aligned}$$

and the results follow from integrating over $P(X)$ and $P(X|D = 1)$.

The sample versions of both *ATE* and *ATT* are suggested by a two-step estimator. Again, first estimate the propensity score. Second, use the estimated score to produce sample estimators.

$$\begin{aligned} \hat{\delta}_{ATE} &= \frac{1}{N} \sum_{i=1}^N Y_i \cdot \frac{D_i - \hat{p}(X_i)}{\hat{p}(X_i) \cdot (1 - \hat{p}(X_i))} \\ \hat{\delta}_{ATT} &= \frac{1}{N_T} \sum_{i=1}^N Y_i \cdot \frac{D_i - \hat{p}(X_i)}{1 - \hat{p}(X_i)} \end{aligned}$$

Using our earlier discussion of steps, this is technically the second step.

Let's see how to do this in Stata. I will move in steps because I want to illustrate to you the importance of trimming the data. First, we need to rescale the outcome variable, as the `teffects` command chokes on large values. So:

```
. gen re78_scaled = re78/10000
. cap n teffects ipw (re78_scaled) (treat age agesq
agecube school schoolsq married nodegree black hispanic re74
re75 u74 u75 interaction1, logit), osample(overlap)
. keep if overlap==0
. drop overlap
. cap n teffects ipw (re78_scaled) (treat age agesq
agecube school schoolsq married nodegree black hispanic re74
re75 u74 u75 interaction1, logit), osample(overlap)
. cap drop overlap
```

Notice the estimated ATE: -0.70 . We have to multiply this by 10,000 since we originally scaled it by 10,000 which is $-0.70 \times 10,000 = -7,000$. In words, inverse probability weighting methods found an ATE that was not only negative, but *very* negative. Why? What happened?

Recall what inverse probability weighting is doing. It is weighting treatment and control units according to $\hat{p}(X)$ which is causing the unit to blow up for very small values of the propensity score. Thus, we will need to trim the data. Here we will do a very small trimming to eliminate the mass of values at the far left tail. Crump et al. [2009] develop a principled method for addressing a lack of overlap. A good rule of thumb, they note, is to keep only observations on the

interval [0.1,0.9], but here I will drop the ones with less than 0.05, and leave it to you to explore this in greater detail.

```
. drop if pscore <= 0.05
```

Now let us repeat the analysis and compare our answer both to what we found when we didn't trim, but also what we got with the experimental ATE.

```
. cap n teffects ipw (re78_scaled) (treat age agesq  
agecube school schoolsq married nodegree black hispanic re74  
re75 u74 u75 interaction1, logit), osample(overlap)
```

Here we find an ATE of \$918 which is significant at $p < 0.12$.
Better, but still not exactly correct and not very precise.

An alternative approach to inverse probability weighting is *nearest neighbor matching* on both the propensity score and covariates themselves. The standard matching strategy is *nearest neighbor matching* where you pair each treatment unit i with one or more comparable control group units j , where comparability is measured in terms of distance to the nearest propensity score. This control outcome is then plugged into a matched sample, and then we simple calculate

$$\widehat{ATT} = \frac{1}{N_T} (Y_i - Y_{i(j)})$$

where $Y_{i(j)}$ is the matched control group unit to i . We will focus on the ATT because of the problems with overlap that we discussed earlier.

For this next part, rerun your do file up to the point where you estimated your inverse probability weighting models. We want to go back to our original data before we dropped the low propensity score units as I want to illustrate how nearest neighbor works. Now type in the following command:

```
. teffects psmatch (re78) (treat age agesq agecube school  
schoolsq married nodegree black hispanic re74 re75 u74 u75  
interaction1, logit), atet gen(pstub_cps) nn(3)
```

A few things to note. First, we are re-estimating the propensity score. Notice the command in the second set of parentheses. We are estimating that equation with logit. Second, this is the ATT, not the ATE. The reason being, we have too many near zeroes in the data to find good matches in the treatment group. Finally, we are matching with three nearest neighbors. Nearest neighbors, in other words, will find the three nearest units in the control group, where "nearest" is measured as closest on the propensity score itself. We then average their actual outcome, and match that average outcome to each treatment unit. Once we have that, we subtract each unit's matched control from its treatment value, and then divide by N_T , the number of treatment units. When we do that in Stata, we get an ATT of \$1,407.75 with a $p < 0.05$. Thus it is both relatively precise and

closer in magnitude to what we find with the experiment itself.

Coarsened Exact Matching There are two kinds of matching we've reviewed so far. There's exact matching which matches a treated unit to all of the control units with the same covariate value. But sometimes this is impossible and therefore there are matching discrepancies.

For instance, say that we are matching continuous age and continuous income. The probability we find another person with the exact same value of both is very small if not zero. This leads therefore to mismatching on the covariates which introduces bias.

The second kind of matching we've discussed are approximate matching methods which specify a metric to find control units that are "close" to the treated unit. This requires a distance metric, such as Euclidean, Mahalanobis or the propensity score. All of these can be implemented in Stata's `teffects`.

Iacus et al. [2012] introduced a kind of exact matching called coarsened exact matching. The idea is very simple. It's based on the notion that sometimes it's possible to do exact matching if we coarsen the data enough. Thus, if we coarsen the data, meaning we create categorical variables (e.g., 0-10 year olds, 11-20 year olds, etc.), then oftentimes we can find exact matches. Once we find those matches, we calculate weights based on where a person fits in some strata and those weights are used in a simple weighted regression.

First, we begin with covariates X and make a copy called X^* . Next we coarsen X^* according to user-defined cutpoints or CEM's automatic binning algorithm. For instance schooling becomes less than high school, high school only, some college, college graduate, post college. Then we create one stratum per unique observation of X^* and place each observation in a stratum. Assign these strata to the original and uncoarsened data, X , and drop any observation whose stratum doesn't contain at least one treated and control unit. You then add weights for stratum size and analyze without matching.

But there are tradeoffs. Larger bins mean more coarsening of the data, which results in fewer strata. Fewer strata result in more diverse observations within the same strata and thus higher covariate imbalance. CEM prunes both treatment and control group units, which changes the parameter of interest, but so long as you're transparent about this and up front about it, readers are willing to give you the benefit of the doubt. Just know, though, that you are not estimating the ATE or the ATT when you start pruning (just as you aren't doing so when you trim propensity scores).

The key benefit of CEM is that it is part of a class of matching methods called monotonic imbalance bounding (MIB). MIB methods bound the maximum imbalance in some feature of the empirical

distributions by an ex ante decision by the user. In CEM, this ex ante choice is the coarsening decision. By choosing the coarsening beforehand, users can control the amount of imbalance in the matching solution. It's also much faster.

There are several ways of measuring imbalance, but here we focus on the $L1(f, g)$ measure which is

$$L1(f, g) = \frac{1}{2} \sum_{l_1 \dots l_k} |f_{l_1 \dots l_k} - g_{l_1 \dots l_k}|$$

where f and g record the relative frequencies for the treatment and control group units. Perfect global balance is indicated by $L1 = 0$. Larger values indicate larger imbalance between the groups, with a maximum of $L1 = 1$. Hence the "imbalance bounding" between 0 and 1.

Now let's get to the fun part: estimation. Here's the command in Stata:

```
. ssc install cem
. cem age (10 20 30 40 60) agesq agecube school schoolsq
married nodegree black hispanic re74 re75 u74 u75 interaction1,
treatment(treat)
. reg re78 treat [iweight=cem_weights], robust
```

The estimated ATE is \$2,771.06, which is much larger than our estimated experimental effect. But, this ensured a high degree of balance on the covariates as can be seen from the output from `cem` command itself.

As can be seen from Table 27, the values of $L1$ are close to zero in most cases. The largest it gets is 0.12 for age squared.

Conclusions Matching methods are an important member of the causal inference arsenal. Propensity scores are an excellent tool to check the balance and overlap of covariates. It's an under appreciated diagnostic and one that you might miss if you only ran regressions. There are extensions for more than two treatments, like multinomial models, but we don't cover those here. The propensity score can make groups comparable but only on the variables used to estimate the propensity score in the first place. There is *no* guarantee you are balancing on unobserved covariates. If you know that there are important, unobservable variables, you will need another tool. Randomization for instance ensures that observable and unobservable variables are balanced.

Table 27: Balance in covariates after coarsened exact matching.

Covariate	L1	Mean	Min	25%	50%	75%	Max
age	.08918	.55337	1	1	0	1	0
agesq	.1155	21.351	33	35	0	49	0
agecube	.05263	626.9	817	919	0	1801	0
school	6.0e-16	-2.3e-14	0	0	0	0	0
schoolsq	5.4e-16	-2.8e-13	0	0	0	0	0
married	1.1e-16	-1.1e-16	0	0	0	0	0
nodegree	4.7e-16	-3.3e-16	0	0	0	0	0
black	4.7e-16	-8.9e-16	0	0	0	0	0
hispanic	7.1e-17	-3.1e-17	0	0	0	0	0
re74	.06096	42.399	0	0	0	0	-94.801
re75	.03756	-73.999	0	0	0	-222.85	-545.65
u74	1.9e-16	-2.2e-16	0	0	0	0	0
u75	2.5e-16	-1.1e-16	0	0	0	0	0
interaction1	.06535	425.68	0	0	0	0	-853.21

Regression discontinuity

Over the last twenty years, there has been significant interest in the *regression-discontinuity design* (RDD). Cook [2008] provides a fascinating history of the procedure, dating back to Thistlethwaite and Campbell [1960] and the multiple implementations of it by its originator, Donald Campbell, an educational psychologist. Cook [2008] documents the early years of the procedure involving Campbell and his students, but notes that by the mid-1970s, Campbell was virtually alone in his use of and interest in this design, despite several attempts to promote it. Eventually he moved on to other things. Campbell and his students made several attempts to bring the procedure into broader use, but despite the publicity, it was not widely adopted in either psychology or education.

The earliest appearance of RDD in economics is an unpublished paper [Goldberger, 1972]. But neither this paper, nor Campbell's work, got into the microeconomist's toolkit until the mid-to-late 1990s when papers using RDD started to appear. Two of the first papers in economics to use a form of it were Angrist and Lavy [1999] and Black [1999]. Angrist and Lavy [1999], which we discuss in detail later, studied the effect of class size on pupil achievement using an unusual feature in Israeli public schools that mechanically created smaller classes when the number of students went over a particular threshold. Black [1999] used a kind of RDD approach when she creatively exploited discontinuities at the geographical level created by school district zoning to estimate people's willingness to pay for better schools. Both papers appear to be the first time since Goldberger [1972] that RDD showed back up in the economics literature.

But 1972 to 1999 is a long time without so much as a peep for what is now considered to be one of the most credible research designs in all of causal inference, so what gives?⁹¹ Cook [2008] says that RDD was "waiting for life" during this time. The conditions in empirical microeconomics had to change before microeconomists realized its potential. Most likely, this was both due to the growing influence of the Rubin causal model among labor economists, as

⁹¹ I should say, for the class of observational data designs. Many, though not all, applied economists and econometricians consider the randomized experiment the gold standard for causal inference.

well as the increased availability of large administrative datasets, including their unusual quirks and features.

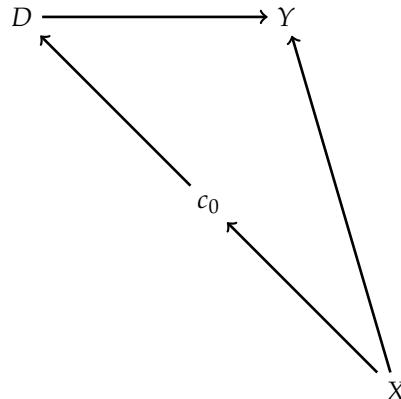
In Thistlewaite and Campbell [1960], the first publication using RDD, the authors studied the effect of merit awards on future academic outcomes. Merit awards were given out to students based on a score, and anyone with a score above some cutoff received the merit award, whereas everyone below that cutoff did not. In their application, the authors knew the mechanism by which the treatment was being assigned to each individual unit – treatment was assigned based on a *cutoff* in some continuous *running variable*. Knowing the treatment assignment allowed them to carefully estimate the causal effect of merit awards on future academic performance.

The reason that RDD was so appealing was because of underlying selection bias. They didn't believe they could simply compare the treatment group (merit award recipients) to the control group (merit award non-recipients), because the two groups were likely very different from one another – on observables, but even more importantly, on unobservables. To use the notation we've been using repeatedly, they did not believe

$$E[Y^0|D = 1] - E[Y^0|D = 0] = 0$$

It was very likely that the recipients were on average of higher overall ability, which directly affects future academic performance. So their solution was to compare only certain students from the treatment and control groups who they thought were *credibly* equivalent – those students who had just high enough scores to get the award and those students with just low enough scores not to get the award.

It's a simple idea, really. Consider the following DAG that illustrates what I mean.



If there is some variable, X , that determines treatment, D , by triggering treatment at c_0 , then isn't this just another form of selection

on observables? If a unit receives treatment because some variable exceeds some threshold, then don't we fully know the treatment assignment? Under what conditions would a comparison of treatment and control group units, incorporating information from the cutoff, yield a credible estimate of the causal effect of treatment?

RDD is appropriate in any situation where a person's entry into the treatment group *jumps* in probability when some running variable, X , exceeds a particular threshold, c_0 . Think about this for a moment: aren't jumps of any kind sort of unnatural? The tendency is for things to change gradually. Charles Darwin once wrote *Natura non facit saltum*, or "nature does not make jumps." Jumps are so unusual that when we see them happen, they beg for some explanation. And in the case of RDD, that "something" is that treatment assignment is occurring based on some running variable, and when that running variable exceeds a particular cutoff value, c_0 , that unit i either gets placed in the treatment group, or that person is *more likely* to be placed in the treatment group. But whichever, the probability of treatment is jumping discontinuously at c_0 .

That's the heart and soul of RDD. We use our knowledge about selection into treatment in order to estimate average treatment effects. More specifically, since we know the probability of treatment assignment changes discontinuously at c_0 , then we will compare people above and below c_0 to estimate a particular kind of average treatment effect called the *local average treatment effect*, or LATE for short [Imbens and Angrist, 1994]. To help make this method concrete, we'll first start out by looking carefully at one of the first papers in economics to use this method [Angrist and Lavy, 1999].

Maimonides Rule and Class Size Krueger [1999] was interested in estimating the causal effect of class size on student test scores using the Tennessee randomized experiment STAR. The same year, another publication came out interested in the same question which used a *natural* experiment [Angrist and Lavy, 1999]. Both students were interested in estimating the causal effect of class size on pupil achievement, but went about the question in very different ways.

One of the earliest references to class size occurs in the Babylonian Talmud, a holy Jewish text completed around the 6th century. One section of the Talmud discusses rules for the determination of class size and pupil-teacher ratios in bible studies. Maimonides was a 12th century Rabbinic scholar who interpreted the Talmud's discussion of class size in what is now known as Maimonides' Rule:

"Twenty-five children may be put in charge of one teacher. If the number in the class exceeds twenty-five, but is not more than forty, two teachers must be appointed."

So what? What does a 12th century Rabbi's interpretation of a 6th century text have to do with causal inference? Because "since 1969, [Maimonides' Rule] has been used to determine the division of enrollment cohorts into classes in Israeli public schools" [Angrist and Lavy, 1999].

The problem with studying the effect of class size on pupil achievement is that class size is likely correlated with the unobserved determinants of pupil achievement too. As a result, any correlation we find is likely biased, and that bias may be large. It may even dominate most of the correlation we find in the first place, making the correlation practically worthless for policy purposes. Those unobservables might include poverty, affluence, enthusiasm/skepticism about the value of education, special needs of students for remedial or advanced instruction, obscure and barely intelligible obsessions of bureaucracies, and so on. Each of these things both determines class size and clouds the effect of class size on pupil achievement because each is independently correlated with pupil achievement.⁹²

However, if adherence to Maimonides' Rule is perfectly rigid, then what would separate a school with a single class of size 40 from the same school with two classes whose average size is 20.5?⁹³ The only difference between them would be the enrollment of a single student. In other words, that one additional student is causing the splitting off of the classes into smaller class sizes. But the two classes should be basically equivalent otherwise. Maimonides' Rule, they argue, appears to be creating exogenous variation in class size.

It turns out Maimonides' Rule has the largest impact on a school with about 40 students in a grade cohort. With cohorts of size 40, 80 and 120 students, the steps down in average class size required by Maimonides' Rule when an additional student enrolls are from 40 to 20.5 ($\frac{41}{2}$), 40 to 27 ($\frac{81}{3}$) and 40 to 30.25 ($\frac{121}{4}$).⁹⁴

Their pupil achievement data are test scores from a short-lived national testing program in Israeli elementary schools. Achievement tests were given in June 1991 and 1992, near the end of the school year, to measure math and reading skills. Average math and reading test scores were rescaled to be on a 100-point scale. The authors then linked this data on test scores with other administrative data on class size and school characteristics.⁹⁵ The unit of observation in the linked data sets is the class and includes data on average test scores in each class, spring class size, beginning-of-year enrollment for each school and grade, a town identifier, school-level index of student SES called "percent disadvantaged" and variables identifying the ethnic and religious composition of the school. Their study was limited to Jewish public schools which account for the vast majority of school children in Israel.

⁹² Put another way, $(Y^1, Y^0) \perp\!\!\!\perp D$ likely does not hold, because D is correlated with the underlying potential outcomes.

⁹³ $\frac{41}{2} = 20.5$.

⁹⁴ Schools also use the percent disadvantaged in a school to allocate supplementary hours of instruction and other school resources which is why Angrist and Lavy [1999] control for it in their regressions.

⁹⁵ This has become increasingly common as administrative data has become digitized and personal computers have become more powerful.

TABLE I
UNWEIGHTED DESCRIPTIVE STATISTICS

Variable	Mean	S.D.	Quantiles						
			0.10	0.25	0.50	0.75	0.90		
A. Full sample									
5th grade (2019 classes, 1002 schools, tested in 1991)									
Class size	29.9	6.5	21	26	31	35	38		
Enrollment	77.7	38.8	31	50	72	100	128		
Percent disadvantaged	14.1	13.5	2	4	10	20	35		
Reading size	27.3	6.6	19	23	28	32	36		
Math size	27.7	6.6	19	23	28	33	36		
Average verbal	74.4	7.7	64.2	69.9	75.4	79.8	83.3		
Average math	67.3	9.6	54.8	61.1	67.8	74.1	79.4		

Figure 24 shows the mean, standard deviation and quantile values for seven variables for the 5th grade across 1,002 schools (from 1991). As can be seen, the mean class size is almost 30 students.

B. ± 5 Discontinuity sample (enrollment 36–45, 76–85, 116–124)

	5th grade		4th grade		3rd grade	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
	(471 classes, 224 schools)		(415 classes, 195 schools)		(441 classes, 206 schools)	
Class size	30.8	7.4	31.1	7.2	30.6	7.4
Enrollment	76.4	29.5	78.5	30.0	75.7	28.2
Percent disadvantaged	13.6	13.2	12.9	12.3	14.5	14.6
Reading size	28.1	7.3	28.3	7.7	24.6	6.2
Math size	28.5	7.4	28.7	7.7	24.8	6.3
Average verbal	74.5	8.2	72.5	7.8	86.2	6.3
Average math	67.0	10.2	68.7	9.1	84.2	7.0

Figure 24: Angrist and Lavy [1999] descriptive statistics

Figure 25: Angrist and Lavy [1999] descriptive statistics for the discontinuity sample.

Angrist and Lavy [1999] present descriptive statistics for what they call a discontinuity sample which is a sample defined as only schools with enrollments greater or less than 5 students: 36,45; 76,85; and 116,125. Average class size is a bit larger in this discontinuity sample than in the overall sample but otherwise very similar to the full sample (see Figure 25).

Papers like these have to figure out how to model the underlying running variable that determines treatment, and in some cases that can be complicated. This is one of those cases. The authors attempt to capture the fact that Maimonides' Rule allows enrollment cohorts of 1-40 to be grouped in a single class, but enrollment cohorts of 41-80 are split into two classes of average size 20.5-40 Enrollment cohorts of 81-120 are split into three classes of average size 27-40 and so on.

Their class size equation is

$$f_{sc} = \frac{e_s}{\text{int}\frac{e_s-1}{40} + 1}$$

where e_s is the beginning-of-year enrollment in school s in a given grade (e.g., 5th grade); f_{sc} is class size assigned to class c in school s for that grade; $\text{int}(n)$ is the largest integer less than or equal to n . They call this the *class size function*. Although the class size function is fixed within schools, in practice enrollment cohorts are not necessarily divided into classes of equal size. But, even though the actual relationship between class size and enrollment size involves many factors, in Israel it clearly has a lot to do with f_{sc} . The authors show this by laying on top of one another f_{sc} from Maimonides' Rule and actual class sizes (Figure 26). Notice the very strong correlation between the two.

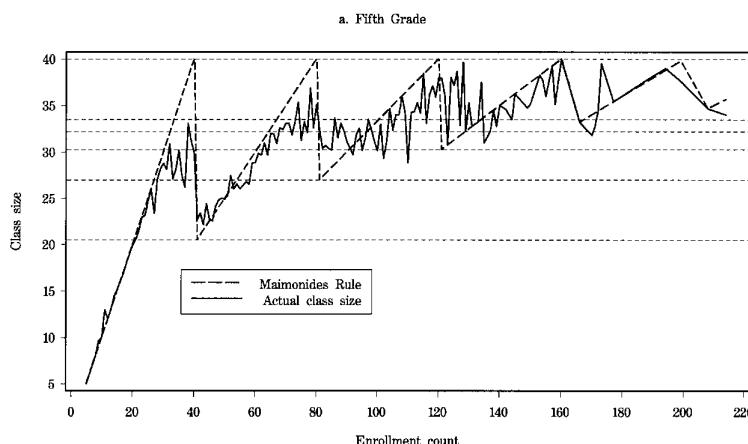


Figure 26: Maimonides' Rule vs. actual class size [Angrist and Lavy, 1999].

Before moving on, look at how great this graph is. The identification strategy told in one picture. Angrist made some really great graphs. Good graphs tell the story. It's worth your time trying to figure out a figure that really conveys your main results or your identification strategy. Readers would prefer it. Okay back to business.

The class size function, f_{sc} , is a mechanical representation of Maimonides' Rule and is highly correlated with actual class size. But it's also highly correlated with average test scores of the fourth and fifth graders. The following picture plots average reading test scores and average values of f_{sc} by enrollment size in enrollment intervals of ten for fifth graders (Figure 27). The figure shows that test scores are generally higher in schools with larger enrollments and larger predicted class sizes, but it also shows an up-and-down pattern

in which average scores by enrollment size mirror the class-size function.

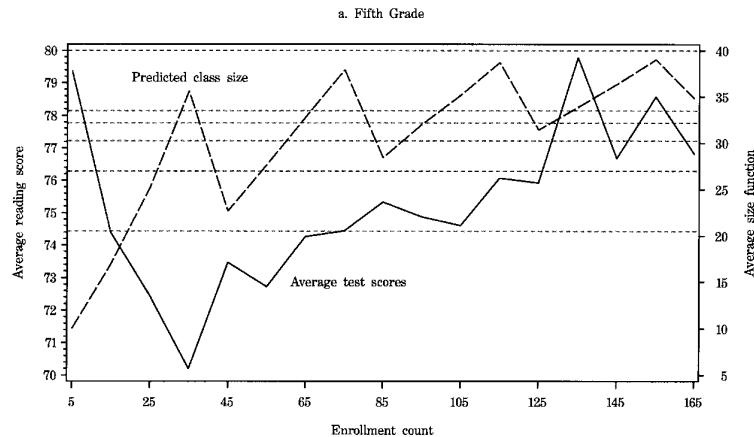


Figure 27: Average reading scores vs. enrollment size [Angrist and Lavy, 1999].

The overall positive correlation between test scores and enrollment is partly attributable to larger schools in Israel being geographically concentrated in larger, more affluent cities. Smaller schools are in poorer “developmental towns” outside the major urban centers. Angrist and Lavy [1999] note that the enrollment size and the percent disadvantaged index measuring the proportion of students from disadvantaged backgrounds are negatively correlated. They control for the “trend” association between test scores and enrollment size and plot the residuals from regressions of average scores and the average of f_{sc} on average enrollment and the percent disadvantaged index for each interval. The estimates for fifth graders imply a reduction in predicted class size of ten students is associated with a 2.2 point increase in average reading scores – a little more than one-quarter of a standard deviation in the distribution of class averages. See the following figures showing the correlation between score residuals and the class size function by enrollment.

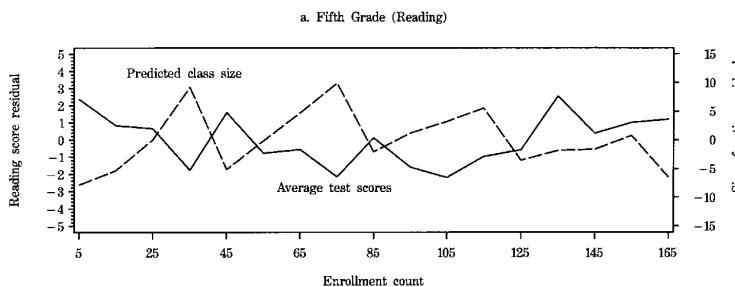


Figure 28: Reading score residual and class size function by enrollment count [Angrist and Lavy, 1999].

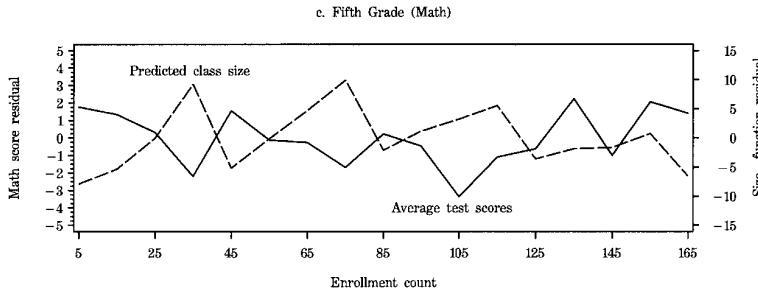


Figure 29: Math score residual and class size function by enrollment count [Angrist and Lavy, 1999].

The visual evidence is strong that class size causes test scores to decrease.⁹⁶ Next, Angrist and Lavy [1999] estimate regression models of the following form

$$y_{isc} = \beta X_s + \delta n_{sc} + \eta_s \mu_c + \varepsilon_{sc}$$

where y_{isc} is pupil i 's score, X_s is a vector of school characteristics, sometimes including functions of enrollment, and n_{sc} is the size of class c in school s . The μ_c is an identical and independently distributed class component, and the term η_s is an identical and independently distributed school component. The class-size coefficient, δ , is the primary parameter of interest.

This equation describes the average potential outcomes of students under alternative assignments of n_{sc} controlling for any effects of X_s . If n_{sc} was randomly assigned conditional on X_s , then δ would be the weighted average response to random variation in class size along the length of the individual causal response functions connecting class size and pupil scores. But n_{sc} is not randomly assigned. Therefore in practice, it is likely correlated with potential outcomes – in this case, the error components in the equation. Estimates of this OLS model are contained in Figure 30.

Though OLS may not have a causal interpretation, using RDD might. The authors go about estimating an RDD model in multiple steps. In the first stage, they estimate the following model:

$$n_{sc} = \pi_0 X_s + \pi f_{sc} + \psi_{sc}$$

where π_j are parameters and the error term is defined as the residual from the population regression of n_{sc} onto X_s and f_{sc} and captures other things that are associated with enrollment. Results from this first stage regressions are presented in Figures 31.

In the second step, the authors calculate the fitted values from the first regression, \hat{n}_{sc} and then estimate the following regression model

$$\bar{y}_{sc} = \beta X_s + \delta \hat{n}_{sc} + \eta_s + [\mu_c + \bar{\varepsilon}_{sc}]$$

⁹⁶ As we will see, graphical evidence is very common in RDD.

TABLE II
OLS ESTIMATES FOR 1991

	5th Grade					
	Reading comprehension			Math		
	(1)	(2)	(3)	(4)	(5)	(6)
<i>Mean score</i>	74.3				67.3	
(<i>s.d.</i>)	(8.1)				(9.9)	
<i>Regressors</i>						
Class size	.221	-.031	-.025	.322	.076	.019
	(.031)	(.026)	(.031)	(.039)	(.036)	(.044)
Percent disadvantaged		-.350	-.351		-.340	-.332
		(.012)	(.013)		(.018)	(.018)
Enrollment			-.002		.017	
			(.006)		(.009)	
Root MSE	7.54	6.10	6.10	9.36	8.32	8.30
R ²	.036	.369	.369	.048	.249	.252
N		2,019			2,018	

Figure 30: OLS regressions [Angrist and Lavy, 1999].

	5th Graders					
	Class size		Reading comprehension		Math	
	(1)	(2)	(3)	(4)	(5)	(6)
A. Full sample						
<i>Means</i>	29.9				74.4	
(<i>s.d.</i>)	(6.5)				(7.7)	
<i>Regressors</i>						
f _{sc}	.704	.542	-.111	-.149	-.009	-.124
	(.022)	(.027)	(.028)	(.035)	(.039)	(.049)
Percent disadvantaged	-.076	-.053	-.360	-.355	-.354	-.338
	(.010)	(.009)	(.012)	(.013)	(.017)	(.018)
Enrollment		.043		.010		.031
		(.005)		(.006)		(.009)
Root MSE	4.56	4.38	6.07	6.07	8.33	8.28
R ²	.516	.553	.375	.377	.247	.255
N		2,019		2,019		2,018
B. Discontinuity sample						
<i>Means</i>	30.8				74.5	
(<i>s.d.</i>)	(7.4)				(8.2)	
<i>Regressors</i>						
f _{sc}	.481	.346	-.197	-.202	-.089	-.154
	(.053)	(.052)	(.050)	(.054)	(.071)	(.077)
Percent disadvantaged	-.130	-.067	-.424	-.422	-.435	-.405
	(.029)	(.028)	(.027)	(.029)	(.039)	(.042)
Enrollment		.086		.003		.041
		(.015)		(.015)		(.022)
Root MSE	5.95	5.58	6.24	6.24	8.58	8.53
R ²	.360	.437	.421	.421	.296	.305
N		471		471		471

Figure 31: First stage regression [Angrist and Lavy, 1999].

	Reading comprehension						Math					
	Full sample			+/- 5 Discontinuity sample			Full sample			+/- 5 Discontinuity sample		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Mean score (s.d.)		74.4 (7.7)			74.5 (8.2)			67.3 (9.6)		67.0 (10.2)		
Regressors												
Class size	-.158 (.040)	-.275 (.066)	-.260 (.081)	-.186 (.104)	-.410 (.113)	-.582 (.181)	-.013 (.056)	-.230 (.092)	-.261 (.113)	-.202 (.131)	-.185 (.151)	-.443 (.236)
Percent disadvantaged	-.372 (.014)	-.369 (.014)	-.369 (.013)		-.477 (.037)	-.461 (.037)	-.355 (.019)	-.350 (.019)	-.350 (.019)	-.459 (.049)	-.459 (.049)	-.435 (.049)
Enrollment		.022 (.009)	.012 (.026)			.053 (.028)		.041 (.012)	.062 (.037)		.079 (.036)	
Enrollment squared/100		.005 (.011)						-.010 (.016)				
Piecewise linear trend				.136 (.032)					.193 (.040)			
Root MSE	6.15	6.23	6.22	7.71	6.79	7.15	8.34	8.40	8.42	9.49	8.79	9.10
N	2019		1961		471		2018		1960		471	

Figure 32: Second stage regressions [Angrist and Lavy, 1999].

Results from this second stage regression are presented in Figure 32.

Compare these second stage regressions to the OLS regressions from earlier (Figure 30). The second stage regressions are all negative and larger in magnitude.

Pulling back for a moment, we can take these results and compare them to what Krueger [1999] found in the Tennessee STAR experiment. Krueger [1999] found effect sizes of around 0.13 - 0.2 standard deviations among pupils and about 0.32 - 0.66 standard deviations in the distribution of class means. Angrist and Lavy [1999] compare their results by calculating the effect size associated with reducing class size by eight pupils (same as STAR). They then multiple this number times their second step estimate for reading scores for fifth graders (-0.275) which gives them an effect size of around 2.2 points or 0.29 standard deviation. Their estimates of effect size for fifth graders are at the low end of the range that Krueger [1999] found in the Tennessee experiment.

Observational are often confounded by a failure to isolate a credible source of exogenous variation in school inputs which leads some researchers to conclude that school inputs don't matter in pupil achievement. But RDD overcomes problems of confounding by exploiting exogenous variation created by *administrative rules*, and as with the STAR experiment, shows that smaller classes appear beneficial to student academic achievement.

Data requirements for RDD RDD is all about finding "jumps" in the probability of treatment as we move along some running variable X. So where do we find these jumps? Where do we find these *discontinuities*? The answer is that humans often embed jumps into rules. Sometimes these embedded rules give us a designs for a careful observational study.

The validity of an RDD doesn't require that the assignment rule

be arbitrary. It only requires that it be known, precise and free of manipulation. The most effective RDD studies involve programs where X has a “hair trigger” that is not tightly related to the outcome being studied. Examples the probability of being arrested for DWI jumping at > 0.08 [Hansen, 2015]; the probability of receiving health-care insurance jumping at age 65 [Card et al., 2008]; the probability of receiving medical attention jumping when birthweight falls below 1,500 grams [Almond et al., 2010]; the probability of attending summer school when grades fall below some minimum level [Jacob and Lefgen, 2004].

In all these kinds of studies, we need data. But specifically, we need a lot of data *around* the discontinuities which itself implies that the datasets useful for RDD are likely very large. In fact, large sample sizes are characteristic features of the RDD. This is also because in the face of strong trends, one typically needs a lot of data. Researchers are typically using administrative data or settings such as birth records where there are many observations.

Definition There are generally accepted two kinds of RDD studies. There are designs where the probability of treatment goes from 0 to 1 at the cutoff, or what is called a “sharp” design. And there are designs where the probability of treatment discontinuously increases at the cutoff. These are often called “fuzzy” designs. In all of these, though, there is some running variable X that upon reaching a cutoff c_0 the likelihood of being in treatment group switches. van der Klaauw [2002] presents the following diagram showing the difference between the two designs:

Sharp RDD is where treatment is a deterministic function of the running variable X .⁹⁷ An example might be Medicare enrollment which happens sharply at age 65 including disability situations. A fuzzy RDD represents a discontinuous “jump” in the probability of treatment when $X > c_0$. In these fuzzy designs, the cutoff is used as an instrumental variable for treatment, like [Angrist and Lavy, 1999] who instrument for class size with the class size function.

More formally, in a sharp RDD, treatment status is a deterministic and discontinuous function of a running variable X_i where

$$D_i = \begin{cases} 1 & \text{if } X_i \geq c_0 \\ 0 & \text{if } X_i < c_0 \end{cases}$$

where c_0 is a known threshold or cutoff. In other words, if you know the value of X_i for unit i , then you know treatment assignment for unit i with certainty. For this reason, people ordinarily think of RDD as a selection on observables observational study.

⁹⁷ Figure 33 calls the running variable “selection variable”. This is because van der Klaauw [2002] is an early paper in the new literature, and the terminology hadn’t yet been hammered out. But they are the same thing.

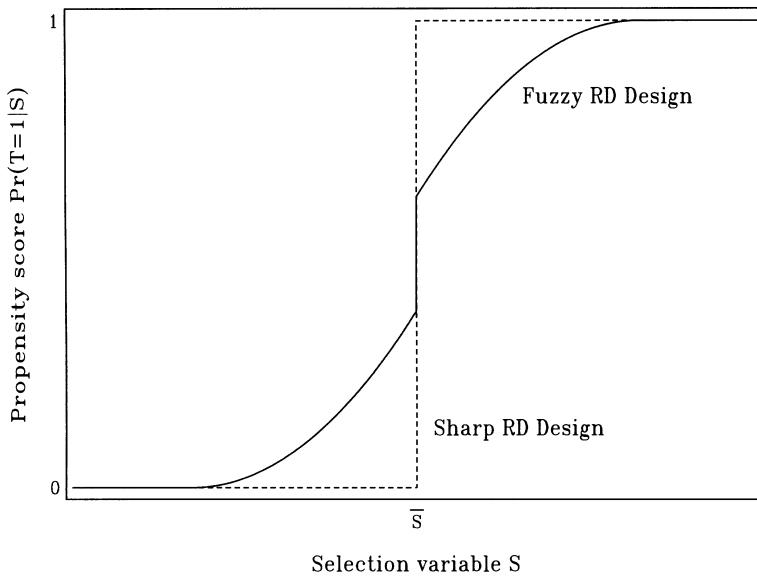


FIGURE 2
ASSIGNMENT IN THE SHARP (DASHED) AND FUZZY (SOLID) RD DESIGN

If we assume constant treatment effects, then in potential outcomes terms, we get

$$\begin{aligned} Y_i^0 &= \alpha + \beta X_i \\ Y_i^1 &= Y_i^0 + \delta \end{aligned}$$

Using the switching equation we get

$$\begin{aligned} Y_i &= Y_i^0 + (Y_i^1 - Y_i^0) D_i \\ Y_i &= \alpha + \beta X_i + \delta D_i + \varepsilon_i \end{aligned}$$

where the treatment effect parameter, δ , is the discontinuity in the conditional expectation function:

$$\begin{aligned} \delta &= \lim_{X_i \rightarrow X_0} E[Y_i^1 | X_i = X_0] - \lim_{X_0 \leftarrow X_i} E[Y_i^0 | X_i = X_0] \\ &= \lim_{X_i \rightarrow X_0} E[Y_i | X_i = X_0] - \lim_{X_0 \leftarrow X_i} E[Y_i | X_i = X_0] \end{aligned}$$

The sharp RDD estimation is interpreted as an average causal effect of the treatment at the discontinuity, which is a kind of local average treatment effect (LATE).

$$\delta_{SRD} = E[Y_i^1 - Y_i^0 | X_i = X_0]$$

Notice the role that *extrapolation* plays in estimating treatment effects with sharp RDD. If unit i is just below c_0 , the $D_i = 0$. But if unit i is just above c_0 , then the $D_i = 1$. See Figure 34.

Figure 33: Sharp vs. Fuzzy RDD [van der Klaauw, 2002].

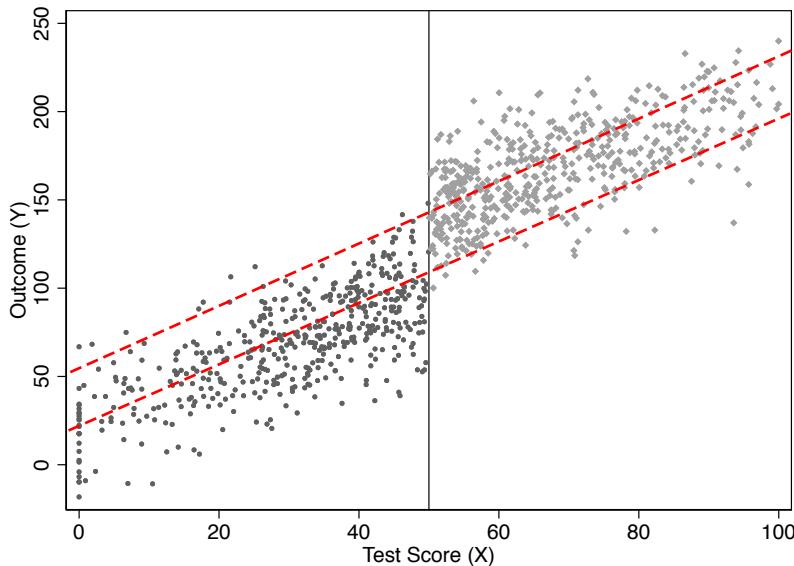


Figure 34: Dashed lines are extrapolations

The key identifying assumption in an RDD is called the continuity assumption. It states

$$E[Y_i^0 | X = c_0] \text{ and } E[Y_i^1 | X = c_0]$$

are continuous (smooth) in X at c_0 . In words, this means that population average potential outcomes, Y^0 and Y^1 , are continuous functions of X at the cutoff, c_0 . That is, the continuity assumption requires that the expected potential outcomes remain continuous through c_0 . Absent the treatment, in other words, the expected potential outcomes wouldn't have jumped; they would've remained smooth functions of X . This implies that all other unobserved determinants of Y are continuously related to the running variable X . Such an assumption should remind you of omitted variable bias. Does there exist some omitted variable wherein the outcome would jump at c_0 even if we disregarded the treatment altogether? If so, then the continuity assumption is violated and our methods do not require the LATE.

Sometimes these abstract ideas become much easier to understand with data, so here is an example of what we mean using a simulation.

```
/// --- Examples using simulated data
. clear
. capture log close
```

```

. set obs 1000
. set seed 1234567

. * Generate running variable
. gen x = rnormal(50, 25)
. replace x=0 if x < 0
. drop if x > 100
. sum x, det

. * Set the cutoff at X=50. Treated if X > 50

. gen D = 0
. replace D = 1 if x > 50
. gen y1 = 25 + 0*D + 1.5*x + rnormal(0, 20)

. twoway (scatter y1 x if D==0, msize(vsmall) msymbol(circle_hollow)) //
(scatter y1 x if D==1, sort mcolor(blue) msize(vsmall) msymbol(circle_hollow)) //
(lfit y1 x if D==0, lcolor(red) msize(small) lwidth(medthin) lpattern(solid)) //
(lfit y1 x, lcolor(dknavy) msize(small) lwidth(medthin) lpattern(solid)), //
xtitle(Test score (X)) xline(50) legend(off)

```

Figure 35 shows the results from this simulation. Notice that the value of Y is changing continuously over X and through c_0 . This is an example of the continuity assumption. It means *absent the treatment itself*, the potential outcomes would've remained a smooth function of X . It is therefore *only* the treatment, triggered at c_0 , that causes the jump. It is worth noting here, as we have in the past, that technically speaking the continuity assumption is not testable because it is based on counterfactuals as so many other identifying assumptions we've reviewed are.

Next we look at an example of discontinuities using simulated data (Figure 36).

```

. gen y = 25 + 40*D + 1.5*x + rnormal(0, 20)
. scatter y x if D==0, msize(vsmall) || scatter y x if D==1, msize(vsmall) legend(off) //
xline(50, lstyle(foreground)) || lfit y x if D ==0, color(red) || //
lfit y x if D ==1, color(red) ytitle("Outcome (Y)") xtitle("Test Score (X)")

```

Notice the jump at the discontinuity in the outcome.

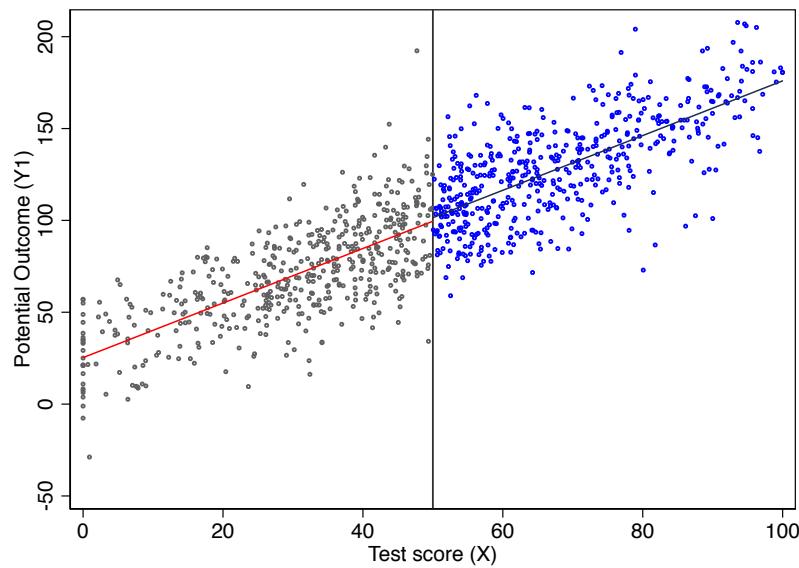


Figure 35: Display of observations from simulation.

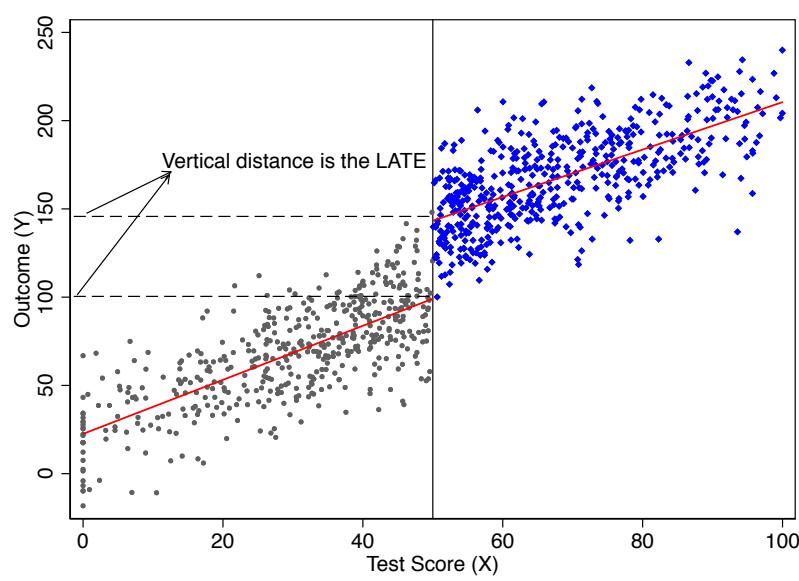


Figure 36: Display of observations discontinuity simulation.

Implementation It is common for authors to transform the running variable X by re-centering at c_0 :

$$Y_i = \alpha + \beta(X_i - c_0) + \delta D_i + \varepsilon_i$$

This doesn't change the interpretation of the treatment effect – only the interpretation of the intercept. Let's use Card et al. [2008] as an example. Medicare is triggered when a person turns 65. So re-center the running variable (age) by subtracting 65:

$$\begin{aligned} Y &= \beta_0 + \beta_1(Age - 65) + \beta_2 Edu \\ &= \beta_0 + \beta_1 Age - \beta_1 65 + \beta_2 Edu \\ &= (\beta_0 - \beta_1 65) + \beta_1 Age + \beta_2 Edu \\ &= \alpha + \beta_1 Age + \beta_2 Edu \end{aligned}$$

where $\alpha = \beta_0 + \beta_1 65$. All other coefficients, notice, have the same interpretation except for the intercept.

Another practical question is nonlinearity. Because sometimes we are fitting local linear regressions around the cutoff, we will pick up an effect because of the imposed linearity if the underlying data generating process is nonlinear. Here's an example from Figure 37:

```
capture drop y
gen x2=x^2
gen x3=x^3

gen y = 25 + 0*D + 2*x + x2 + rnormal(0, 20)
scatter y x if D==0, msize(vsmall) || scatter y x if D==1, msize(vsmall) legend(off) ||
xline(50, lstyle(foreground)) ytitle("Outcome (Y)") xtitle("Test Score (X)")
```

In this situation, we would need some way to model the nonlinearity below and above the cutoff to check whether, even given the nonlinearity, there had been a jump in the outcome at the discontinuity.

Suppose that the nonlinear relationships is

$$E[Y_i^0 | X_i] = f(X_i)$$

for some reasonably smooth function $f(X_i)$. In that case, we'd fit the regression model:

$$Y_i = f(X_i) + \delta D_i + \eta_i$$

Since $f(X_i)$ is counterfactual for values of $X_i > c_0$, how will we model the nonlinearity? There are two ways of approximating $f(X_i)$. First,

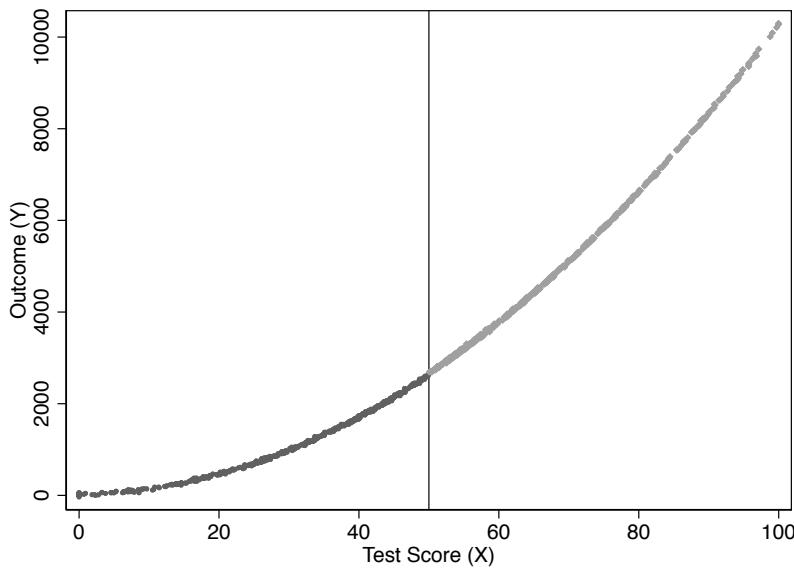


Figure 37: Simulated nonlinear data from Stata

let $f(X_i)$ equal a p^{th} order polynomial:

$$Y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_p x_i^p + \delta D_i + \eta_i$$

This approach, though, has recently been found to introduce bias [Gelman and Imbens, 2016]. Those authors recommend using local linear regressions with linear and quadratic forms only. Another way of approximating $f(X_i)$ is to use a nonparametric kernel, which I will discuss later.

But let's stick with this example where we are using p^{th} order polynomials, just so you know the history of this method and understand better what is being done. We can generate this function, $f(X_i)$, by allowing the x_i terms to differ on both sides of the cutoff by including them both individually and interacting them with D_i . In that case, we have:

$$\begin{aligned} E[Y_i^0 | X_i] &= \alpha + \beta_{01} \tilde{X}_i + \cdots + \beta_{0p} \tilde{X}_i^p \\ E[Y_i^1 | X_i] &= \alpha + \beta_{11} \tilde{X}_i + \cdots + \beta_{1p} \tilde{X}_i^p \end{aligned}$$

where \tilde{X}_i is the re-centered running variable (i.e., $X_i - c_0$). Centering at c_0 ensures that the treatment effect at $X_i = X_0$ is the coefficient on D_i in a regression model with interaction terms. As Lee and Lemieux [2010] note, allowing different functions on both sides of the discontinuity should be the main results in an RDD paper.

To derive a regression model, first note that the observed values must be used in place of the potential outcomes

$$E[Y|X]E[Y^0|X] + (E[Y^1|X] - E[Y^0|X])D$$

Your regression model then is

$$\begin{aligned} Y &= \alpha + \beta_{01}\tilde{x}_i + \dots + \beta_{0p}\tilde{x}_i^p + \delta D_i \\ &\quad + \beta_1^*\tilde{x}_i + \dots + \beta_p^*D_i\tilde{x}_i^p + \varepsilon_i \end{aligned}$$

where $\beta_1^* = \beta_{11} - \beta_{01}$, and $\beta_p^* = \beta_{1p} - \beta_{0p}$. The equation we looked at earlier was just a special case of the above equation with $\beta_1^* = \beta_p^* = 0$. The treatment effect at c_0 is δ . And the treatment effect at $X_i - c_0 > 0$ is $\delta + \beta_1^*c + \dots + \beta_p^*c^p$.

```
. capture drop y x2 x3
. gen x2 = x*x
. gen x3 = x*x*x
. gen y = 10000 + 0*D - 100*x +x2 + rnormal(0, 1000)
. reg y D x x2 x3
. predict yhat

. scatter y x if D==0, msize(vsmall) || scatter y x if D==1,
msize(vsmall) legend(off) xline(140, lstyle(foreground)) ylabel(none) ||
line yhat x if D ==0, color(red) sort || line yhat x if D ==1, sort color(red)
xtitle("Test Score (X)") ytitle("Outcome (Y)")
```

But, as we mentioned earlier, [Gelman and Imbens \[2016\]](#) has recently discouraged the use of higher order polynomials when estimating local linear regressions. The alternative is to use kernel regression. The nonparametric kernel method has problems because you are trying to estimate regressions at the cutoff point which can result in a boundary problem (see Figure 38 from [Hahn et al. \[2001\]](#)).

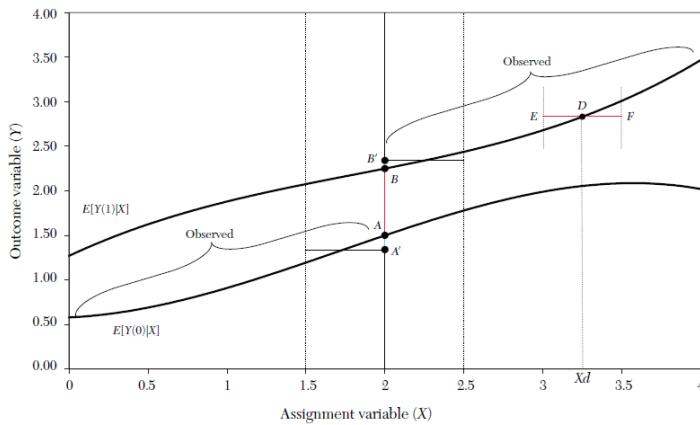


Figure 38: Illustration of a boundary problem

While the true effect in this diagram is AB , with a certain bandwidth a rectangular kernel would estimate the effect as $A'B'$, which

is as you can see a biased estimator. There is systematic bias with the kernel method if the underlying nonlinear function, $f(X)$, is upwards or downwards sloping.

The standard solution to this problem is to run local linear non-parametric regression [Hahn et al., 2001]. In the case described above, this would substantially reduce the bias. So what is that? Think of kernel regression as a weighted regression restricted to a window (hence “local”). The kernel provides the weights to that regression. Stata’s poly command estimates kernel-weighted local polynomial regression. A rectangular kernel would give the same result as taking $E[Y]$ at a given bin on X . The triangular kernel gives more importance to the observations closest to the center.

The model is some version of:

$$(\hat{a}, \hat{b}) =_{a,b} \sum_{i=1}^n (y_i - a - b(x_i - c_0))^2 K\left(\frac{x_i - c_0}{h}\right) 1(x_i > c_0) \quad (79)$$

While estimating this in a given window of width h around the cutoff is straightforward, what’s not straightforward is knowing how large or small to make the window.⁹⁸ So this method is sensitive to the choice of bandwidth. Optimal bandwidth selection has become available [Imbens and Kalyanaraman, 2011].

Card et al. [2008] Card et al. [2008] is an example of a sharp RDD, because it focuses on the provision of universal healthcare insurance for the elderly – Medicare at age 65. What makes this a policy-relevant question is because questions regarding universal insurance have become highly relevant because of the debates surrounding the Affordable Care Act. But also because of the sheer size of Medicare. In 2014, Medicare was 14% of the federal budget at \$505 billion.

Approximately 20% of non elderly adults in the US lacked insurance in 2005. Most were from lower-income families, and nearly half were African American or Hispanic. Many analysts have argued that unequal insurance coverage contributes to disparities in health care utilization and health outcomes across socioeconomic status. But, even among the insured, there is heterogeneity in the form of different copays, deductibles and other features that affect use. Evidence that better insurance causes better health outcomes is limited because health insurance suffers from deep selection bias. Both supply and demand for insurance depend on health status, confounding observational comparisons between people with different insurance characteristics.

The situation for elderly looks very different, though. Less than 1% of the elderly population are uninsured. Most have fee-for-service Medicare coverage. And that transition to Medicare occurs sharply at

⁹⁸ You’ll also see the window referred to as the bandwidth. They mean the same thing.

age 65 – the threshold for Medicare eligibility.

The authors estimate a reduced form model measuring the causal effect of health insurance status on health care usage:

$$y_{ija} = X_{ija}\alpha + f_k(\alpha; \beta) + \sum_k C_{ija}^k \delta^k + u_{ija}$$

where i indexes individuals, j indexes a socioeconomic group, a indexes age, u_{ija} indexes the unobserved error, y_{ija} health care usage, X_{ija} a set of covariates (e.g., gender and region), $f_j(\alpha; \beta)$ a smooth function representing the age profile of outcome y for group j , and C_{ija}^k ($k = 1, 2, \dots, K$) are characteristics of the insurance coverage held by the individual such as copayment rates. The problem with estimating this model, though, is that insurance coverage is endogenous: $\text{cov}(u, C) \neq 0$. So the authors use as identification of the age threshold for Medicare eligibility at 65, which they argue is credibly exogenous variation in insurance status. See Figure 39 as an example of the correlation between age and insurance status.

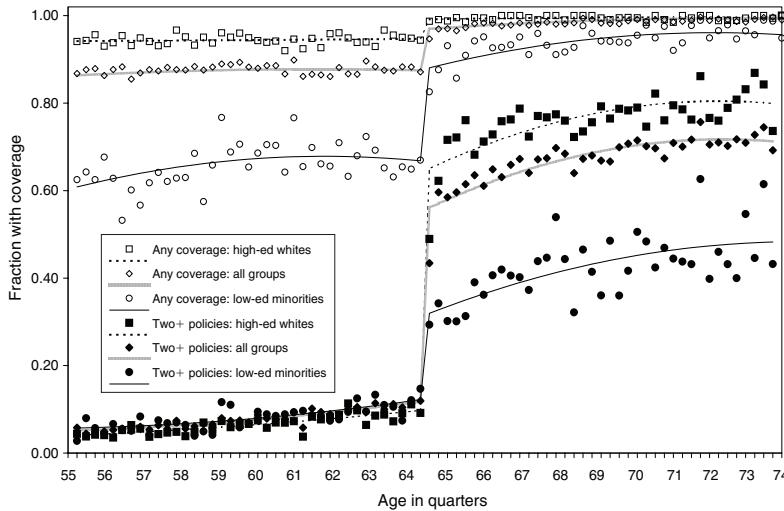


FIGURE 1. COVERAGE BY ANY INSURANCE AND BY TWO OR MORE POLICIES, BY AGE AND DEMOGRAPHIC GROUP

Figure 39: Insurance status and age

Suppose health insurance coverage can be summarized by two dummy variables: C_{ija}^1 (any coverage) and C_{ija}^2 (generous insurance). Card et al. [2008] estimate the following linear probability models:

$$\begin{aligned} C_{ija}^1 &= X_{ija}\beta_j^1 + g_j^1(a) + D_a\pi_j^1 + v_{ija}^1 \\ C_{ija}^2 &= X_{ija}\beta_j^2 + g_j^2(a) + D_a\pi_j^2 + v_{ija}^2 \end{aligned}$$

where β_j^1 and β_j^2 are group-specific coefficients, $g_j^1(a)$ and $g_j^2(a)$ are

smooth age profiles for group j , and D_a is a dummy if the respondent is equal to or over age 65. Recall the reduced form model:

$$y_{ija} = X_{ija}\alpha + f_k(\alpha; \beta) + \sum_k C_{ija}^k \delta^k + u_{ija}$$

Combining the C_{ija} equations, and rewriting the reduced form model, we get:

$$y_{ija} = X_{ija} \left(\alpha_j + \beta_j^1 \delta_j^1 + \beta_j^2 \delta_j^2 \right) h_j(a) + D_a \pi_j^y + v_{ija}^y$$

where $h(a) = f_j(a) + \delta^1 g_j^1(a) + \delta^2 g_j^2(a)$ is the reduced form age profile for group j , $\pi_j^y = \pi_j^1 \delta^1 + \pi_j^2 \delta^2$ and $v_{ija}^y = u_{ija} + v_{ija}^1 \delta^1 + v_{ija}^2 \delta^2$ is the error term. Assuming that the profiles $f_j(a)$, $g_j^1(a)$ and $g_j^2(a)$ are continuous at age 65 (i.e., the continuity assumption necessary for identification), then any discontinuity in y is due to insurance. The magnitudes will depend on the size of the insurance changes at age 65 (π_j^1 and π_j^2) and on the associated causal effects (δ^1 and δ^2).

For some basic health care services, such as routine doctor visits, it may be that the only thing that matters is insurance. But, in those situations, the implied discontinuity in Y at age 65 for group j will be proportional to the change in insurance status experienced by that group. For more expensive or elective services, the generosity of the coverage may matter. For instance, if patients are unwilling to cover the required copay or if the managed care program won't cover the service. This creates a potential identification problem in interpreting the discontinuity in y for any one group. Since π_j^y is a linear combination of the discontinuities in coverage and generosity, δ^1 and δ^2 can be estimated by a regression across groups:

$$\pi_j^y = \delta^0 + \delta^1 \pi_j^1 + \delta^2 \pi_j^2 + e_j$$

where e_j is an error term reflecting a combination of the sampling errors in π_j^y , π_j^1 and π_j^2 .

Card et al. [2008] use a couple of different datasets – one a standard survey and the other administrative records from hospitals in three states. First, they use the 1992-2003 National Health Interview Survey (NHIS). The NHIS reports respondents' birth year, birth month, and calendar quarter of the interview. Authors used this to construct an estimate of age in quarters at date of interview. A person who reaches 65 in the interview quarter is coded as age 65 and 0 quarters. Assuming a uniform distribution of interview dates, one-half of these people be 0-6 weeks younger than 65 and one-half will be 0-6 weeks older. Analysis is limited to people between 55 and 75. The final sample has 160,821 observations.

The second dataset is hospital discharge records for California, Florida and New York. These records represent a complete census of discharges from all hospitals in the three states except for federally regulated institutions. The data files include information on age in months at the time of admission. Their sample selection criteria is to drop records for people admitted as transfers from other institutions, and limit people between 60 and 70 years of age at admission. Sample sizes are 4,017,325 (California), 2,793,547 (Florida) and 3,121,721 (New York).

Some institutional details about the Medicare program may be helpful. Medicare is available to people who are at least 65 and have worked 40 quarters or more in covered employment or have a spouse who did. Coverage is available to younger people with severe kidney disease and recipients of Social Security Disability Insurance. Eligible individuals can obtain Medicare hospital insurance (Part A) free of charge, and medical insurance (Part B) for a modest monthly premium. Individuals receive notice of their impending eligibility for Medicare shortly before their 65th birthday and are informed they have to enroll in it and choose whether to accept Part B coverage. Coverage begins on the first day of the month in which they turn 65.

There are five insurance-related variables: probability of Medicare coverage, any health insurance coverage, private coverage, two or more forms of coverage, and individual's primary health insurance is managed care. Data are drawn from the 1999-2003 NHIS and for each characteristic, authors show the incidence rate at ages 63-64 and the change at age 65 based on a version of the C_K equations that include a quadratic in age, fully interacted with a post-65 dummy as well as controls for gender, education, race/ethnicity, region and sample year. Alternative specifications were also used, such as a parametric model fit to a narrower age window (ages 63-67) and a local linear regression specification using a chosen bandwidth. Both show similar estimates of the change at age 65.

The authors present their findings in Table, which is reproduced here as Figure 40. The way that you read this table is that the odd numbered columns show the mean values for comparison group (63-64 year olds) and the even numbered columns show the *average treatment effect* for this population that complies with the treatment. We can see, not surprisingly, that the effect of receiving Medicare is to cause a very large increase of being on Medicare, as well as reducing coverage on private and managed care.

Formal identification in an RDD relating to some outcome (insurance coverage) to a treatment (MEDicare age-eligibility) that itself depends on some running variable, age, relies on the continuity assumptions that we discussed earlier. That is, we must assume that

TABLE 1—INSURANCE CHARACTERISTICS JUST BEFORE AGE 65 AND ESTIMATED DISCONTINUITIES AT AGE 65

	On Medicare		Any insurance		Private coverage		2+ Forms coverage		Managed care	
	Age 63–4 (1)	RD at 65 (2)	Age 63–4 (3)	RD at 65 (4)	Age 63–4 (5)	RD at 65 (6)	Age 63–4 (7)	RD at 65 (8)	Age 63–4 (9)	RD at 65 (10)
Overall sample	12.3	59.7 (4.1)	87.9	9.5 (0.6)	71.8	−2.9 (1.1)	10.8	44.1 (2.8)	59.4	−28.4 (2.1)
<i>Classified by ethnicity and education:</i>										
White non-Hispanic:										
High school dropout	21.1	58.5 (4.6)	84.1	13.0 (2.7)	63.5	−6.2 (3.3)	15.0	44.5 (4.0)	48.1	−25.0 (4.5)
High school graduate	11.4	64.7 (5.0)	92.0	7.6 (0.7)	80.5	−1.9 (1.6)	10.1	51.8 (3.8)	58.9	−30.3 (2.6)
At least some college	6.1	68.4 (4.7)	94.6	4.4 (0.5)	85.6	−2.3 (1.8)	8.8	55.1 (4.0)	69.1	−40.1 (2.6)
Minority:										
High school dropout	19.5	44.5 (3.1)	66.8	21.5 (2.1)	33.2	−1.2 (2.5)	11.4	19.4 (1.9)	39.1	−8.3 (3.1)
High school graduate	16.7	44.6 (4.7)	85.2	8.9 (2.8)	60.9	−5.8 (5.1)	13.6	23.4 (4.8)	54.2	−15.4 (3.5)
At least some college	10.3	52.1 (4.9)	89.1	5.8 (2.0)	73.3	−5.4 (4.3)	11.1	38.4 (3.8)	66.2	−22.3 (7.2)
<i>Classified by ethnicity only:</i>										
White non-Hispanic (all)	10.8	65.2 (4.6)	91.8	7.3 (0.5)	79.7	−2.8 (1.4)	10.4	51.9 (3.5)	61.9	−33.6 (2.3)
Black non-Hispanic (all)	17.9	48.5 (3.6)	84.6	11.9 (2.0)	57.1	−4.2 (2.8)	13.4	27.8 (3.7)	48.2	−13.5 (3.7)
Hispanic (all)	16.0	44.4 (3.7)	70.0	17.3 (3.0)	42.5	−2.0 (1.7)	10.8	21.7 (2.1)	52.9	−12.1 (3.7)

Note: Entries in odd-numbered columns are percentages of age 63–64-year-olds in group with insurance characteristic shown in column heading. Entries in even-numbered columns are estimated regression discontinuities at age 65, from models that include quadratic control for age, fully interacted with dummy for age 65 or older. Other controls include indicators for gender, race/ethnicity, education, region, and sample year. Estimates are based on linear probability models fit to pooled samples of 1999–2003 NHIS.

Figure 40: Card et al. [2008] Table 1

the conditional expectation functions for both potential outcomes is continuous at age=65. This means that both $E[Y^0|a]$ and $E[Y^1|a]$ are continuous through age of 65. If that assumption is plausible, then the average treatment effect at age 65 is identified as:

$$\lim_{a \rightarrow 65^-} E[y^1|a] - \lim_{a \rightarrow 65^+} E[y^0|a]$$

The continuity assumption requires that all other factors, observed and unobserved, that affect insurance coverage are trending smoothly at the cutoff, in other words. But what else changes at age 65 other than Medicare eligibility? Employment changes. Typically, 65 is the traditional age when people retire from the labor force. Any abrupt change in employment could lead to differences in health care utilization if non workers have more time to visit doctors.

The authors need to, therefore, investigate this possible confounder. They do this by testing for any potential discontinuities at age 65 for confounding variables using a third dataset – the March CPS 1996–2004. And they ultimately find no evidence for discontinuities in employment at age 65 (Figure 41).

Next the authors investigate the impact that Medicare had on access to care and utilization using the NHIS data. Since 1997, NHIS has asked four questions. They are:

“During the past 12 months has medical care been delayed for this person because of worry about the cost?”

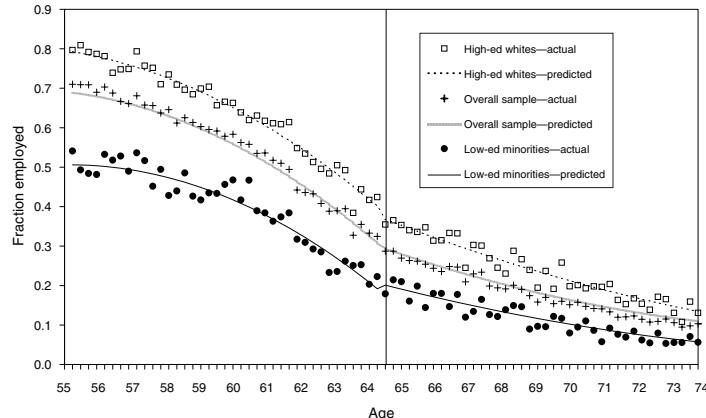


FIGURE 2. EMPLOYMENT RATES BY AGE AND DEMOGRAPHIC GROUP (1992–2003 NHIS)

Figure 41: Investigating the CPS for discontinuities at age 65 [Card et al., 2008]

"During the past 12 months was there any time when this person needed medical care but did not get it because (this person) could not afford it?"

"Did the individual have at least one doctor visit in the past year?"

"Did the individual have one or more overnight hospital stays in the past year?"

Estimates from this analysis are in Figure 42. Again, the odd numbered columns are the baseline, and the even numbered columns are the average treatment effect. Standard errors are in parenthesis below coefficient estimates in the even numbered columns. There's a few encouraging findings from this table. First, the share of the relevant population who delayed care the previous year fell 1.8 points, and similar for the share who did not get care at all in the previous year. The share who saw a doctor went up slightly, as did the share who stayed at a hospital. These are not very large effects in magnitude, it is important to note, but they are relatively precisely estimated. Note that these effects differed considerably by race and ethnicity as well as education.

Having shown modest effects on care and utilization, the authors turn to examining the kinds of care they received by examining specific changes in hospitalizations. Figure 43 shows the effect of Medicare on hip and knee replacements by race. The effects are largest for Whites.

In conclusion, the authors find that universal healthcare coverage for the elderly increases care and utilization, as well as coverage. In a subsequent study [Card et al., 2009], the authors examined the impact of Medicare on mortality and find slight decreases in mortality rates (see Figure 44).

TABLE 3—MEASURES OF ACCESS TO CARE JUST BEFORE 65 AND ESTIMATED DISCONTINUITIES AT 65

	1997–2003 NHIS				1992–2003 NHIS			
	Delayed care last year		Did not get care last year		Saw doctor last year		Hospital stay last year	
	Age 63–64	RD at 65	Age 63–64	RD at 65	Age 63–64	RD at 65	Age 63–64	RD at 65
Overall sample	7.2	-1.8 (0.4)	4.9	-1.3 (0.3)	84.8	1.3 (0.7)	11.8	1.2 (0.4)
<i>Classified by ethnicity and education:</i>								
White non-Hispanic:								
High school dropout	11.6	-1.5 (1.1)	7.9	-0.2 (1.0)	81.7	3.1 (1.3)	14.4	1.6 (1.3)
High school graduate	7.1	0.3 (2.8)	5.5	-1.3 (2.8)	85.1	-0.4 (1.5)	12.0	0.3 (0.7)
At least some college	6.0	-1.5 (0.4)	3.7	-1.4 (0.3)	87.6	0.0 (1.3)	9.8	2.1 (0.7)
Minority:								
High school dropout	13.6	-5.3 (1.0)	11.7	-4.2 (0.9)	80.2	5.0 (2.2)	14.5	0.0 (1.4)
High school graduate	4.3	-3.8 (3.2)	1.2	1.5 (3.7)	84.8	1.9 (2.7)	11.4	1.8 (1.4)
At least some college	5.4	-0.6 (1.1)	4.8	-0.2 (0.8)	85.0	3.7 (3.9)	9.5	0.7 (2.0)
<i>Classified by ethnicity only:</i>								
White non-Hispanic	6.9	-1.6 (0.4)	4.4	-1.2 (0.3)	85.3	0.6 (0.8)	11.6	1.3 (0.5)
Black non-Hispanic (all)	7.3	-1.9 (1.1)	6.4	-0.3 (1.1)	84.2	3.6 (1.9)	14.4	0.5 (1.1)
Hispanic (all)	11.1	-4.9 (0.8)	9.3	-3.8 (0.7)	79.4	8.2 (0.8)	11.8	1.0 (1.6)

Note: Entries in odd numbered columns are mean of variable in column heading among people ages 63–64. Entries in even numbered columns are estimated regression discontinuities at age 65, from models that include linear control for age interacted with dummy for age 65 or older (columns 2 and 4) or quadratic control for age, interacted with dummy for age 65 and older (columns 6 and 8). Other controls in models include indicators for gender, race/ethnicity, education, region, and sample year. Sample in columns 1–4 is pooled 1997–2003 NHIS. Sample in columns 5–8 is pooled 1992–2003 NHIS. Samples for regression models include people ages 55–75 only. Standard errors (in parentheses) are clustered by quarter of age.

Figure 42: Investigating the NHIS for the impact of Medicare on care and utilization [Card et al., 2008]

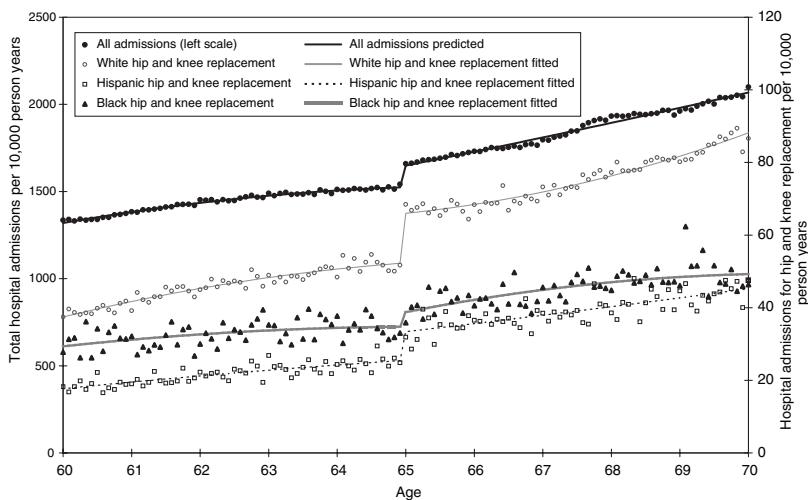


FIGURE 3. HOSPITAL ADMISSION RATES BY RACE/ETHNICITY

Figure 43: Changes in hospitalizations [Card et al., 2008]

TABLE V
REGRESSION DISCONTINUITY ESTIMATES OF CHANGES IN MORTALITY RATES

	7 days	14 days	28 days	90 days	180 days	365 days
<i>Estimated discontinuity at age 65 ($\times 100$)</i>						
Fully interacted quadratic with no additional controls	-1.1 (0.2)	-1.0 (0.2)	-1.1 (0.3)	-1.1 (0.3)	-1.2 (0.4)	-1.0 (0.4)
Fully interacted quadratic plus additional controls	-1.0 (0.2)	-0.8 (0.2)	-0.9 (0.3)	-0.9 (0.3)	-0.8 (0.3)	-0.7 (0.4)
Fully interacted cubic plus additional controls	-0.7 (0.3)	-0.7 (0.2)	-0.6 (0.4)	-0.9 (0.4)	-0.9 (0.5)	-0.4 (0.5)
Local linear regression procedure fit separately to left and right with rule-of-thumb bandwidths	-0.8 (0.2)	-0.8 (0.2)	-0.8 (0.2)	-0.9 (0.2)	-1.1 (0.3)	-0.8 (0.3)
Mean of dependent variable (%)	5.1	7.1	9.8	14.7	18.4	23.0

Notes. Standard errors in parentheses. Dependent variable is indicator for death within interval indicated by column heading. Entries in rows (1)–(3) are estimated coefficients of dummy for age over 65 from models that include a quadratic polynomial in age (rows (1) and (2)) or a cubic polynomial in age (row (3)) fully interacted with a dummy for age over 65. Models in rows (4)–(6) include the following additional controls: a dummy for people who are within 1 month of their 65th birthday, dummies for year, month, sex, race/electicity, and Social Security Administration, and uncentered fixed effects for each hospital. Estimated standard errors in row (4) are estimated by bootstrapping from local linear regression procedure, fit separately to the left and right, with independently selected bandwidths from a rule-of-thumb procedure suggested by Fan and Gijbels (1996). Sample includes 407,386 observations on patients between the ages of 60 and 70 admitted to California hospitals between January 1, 1992, and November 30, 2002, for unplanned admission through the ED who have nonmissing Social Security numbers. All coefficients and their SEs have been multiplied by 100.

We will return to the question of healthcare coverage when we cover the Medicaid Oregon experiment in the instrumental variables chapter, but for now we stop.

Fuzzy RDD In the sharp RDD, treatment was *determined* when $X_i \geq c_0$. But that kind of deterministic assignment does not always happen. Sometimes there is a discontinuity, but it's not entirely deterministic, though it nonetheless is associated with a discontinuity in treatment assignment. When there is an increase in the *probability* of treatment assignment, we have a *fuzzy* RDD. The formal definition of a probabilistic treatment assignment is

$$\lim_{X_i \rightarrow c_0} \Pr(D_i = 1 | X_i = X_0) \neq \lim_{c_0 \leftarrow X_i} \Pr(D_i = 1 | X_i = X_0)$$

In other words, the conditional probability is becoming discontinuous as X approaches c_0 in the limit. A visualization of this is presented from [Imbens and Lemieux \[2008\]](#) in Figure 45:

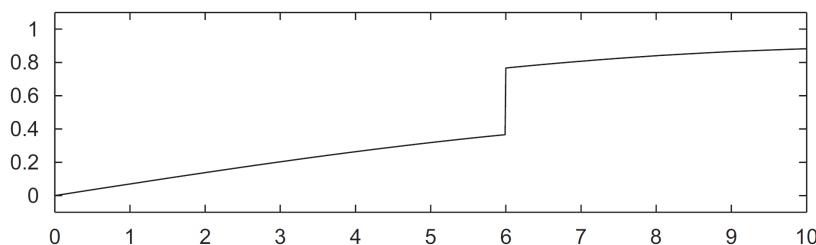


Fig. 3. Assignment probabilities (FRD).

Figure 44: Mortality and Medicare [Card et al., 2009]

Figure 45: [Imbens and Lemieux \[2008\]](#), Figure 3. Horizontal axis is the running variable. Vertical axis is the conditional probability of treatment at each value of the running variable.

As you can see in this picture, the treatment assignment is increasing even before c_0 , but is not fully assigned to treatment above c_0 .

Rather, the fraction of the units in the treatment jumps at c_0 . This is what a fuzzy discontinuity looks like.

The identifying assumptions are the same under fuzzy designs as they are under sharp designs: they are the continuity assumptions. For identification, we must assume that the conditional expectation of the potential outcomes (e.g., $E[Y^0|X < c_0]$) is changing smoothly through c_0 . What changes at c_0 is the treatment assignment probability. An illustration of this identifying assumption is in Figure 46.

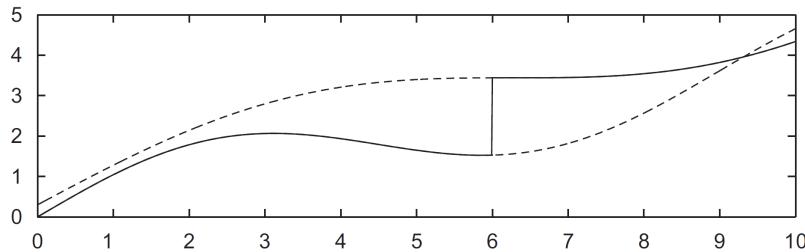


Figure 46: Potential and observed outcome regressions [Imbens and Lemieux, 2008]

Calculating the average treatment effect under a fuzzy RDD is very similar to how we calculate an average treatment effect with instrumental variables. Specifically, it's the ratio of a reduced form difference in mean outcomes around the cutoff and a reduced form difference in mean treatment assignment around the cutoff.

$$\delta_{\text{Fuzzy RDD}} = \frac{\lim_{X \rightarrow X_0} E[Y|X = X_0] - \lim_{X_0 \leftarrow X} E[Y|X = X_0]}{\lim_{X \rightarrow X_0} E[D|X = X_0] - \lim_{X_0 \leftarrow X} E[D|X = X_0]}$$

This can be calculated with software in Stata, such as `ivregress 2sls`. The assumptions for identification are the same as those with instrumental variables: there are caveats about the complier vs. the defier populations, statistical tests (e.g., weak instrument using F tests on the first stage), etc.

One can use both T_i as well as the interaction terms as instruments for the treatment D_i . If one uses only T_i as an instrumental variable, then it is a “just identified” model which usually has good finite sample properties. In the just identified case, the first stage would be:

$$D_i = \gamma_0 + \gamma_1 X_i + \gamma_2 X_i^2 + \dots + \gamma_p X_i^p + \pi T_i + \zeta_{1i}$$

where π is the causal effect of T on the conditional probability of treatment. The fuzzy RDD reduced form is:

$$Y_i = \mu + \kappa_1 X_i + \kappa_2 X_i^2 + \dots + \kappa_p X_i^p + \rho \pi T_i + \zeta_{2i}$$

As in the sharp RDD case, one can allow the smooth function to be different on both sides of the discontinuity. The second stage model

with interaction terms would be the same as before:

$$\begin{aligned} Y_i &= \alpha + \beta_{01}\tilde{x}_i + \beta_{02}\tilde{x}_i^2 + \cdots + \beta_{0p}\tilde{x}_i^p \\ &\quad + \rho D_i + \beta_1^* D_i \tilde{x}_i + \beta_2^* D_i \tilde{x}_i^2 + \cdots + \beta_p^* D_i \tilde{x}_i^p + \eta_i \end{aligned}$$

Where \tilde{x} are now not only normalized with respect to c_0 but are also fitted values obtained from the first stage regressions. Again, one can use both T_i as well as the interaction terms as instruments for D_i . If we only used T , the estimated first stage would be:

$$\begin{aligned} D_i &= \gamma_{00} + \gamma_{01}\tilde{X}_i + \gamma_{02}\tilde{X}_i^2 + \cdots + \gamma_{0p}\tilde{X}_i^p \\ &\quad + \pi T_i + \gamma_1^* \tilde{X}_i T_i + \gamma_2^* \tilde{X}_i^2 T_i + \cdots + \gamma_p^* T_i + \zeta_{1i} \end{aligned}$$

We would also construct analogous first stages for $\tilde{X}_i D_i, \dots, \tilde{X}_i^p D_i$.

As Hahn et al. [2001] point out, one needs the same assumptions for identification as one needs with IV. As with other binary instrumental variables, the fuzzy RDD is estimating the local average treatment effect (LATE) [Imbens and Angrist, 1994], which is the average treatment effect for the compliers. In RDD, the compliers are those whose treatment status changed as we moved the value of x_i from just to the left of c_0 to just to the right of c_0 .

Challenges to identification The identifying assumption for RDD to estimate a causal effect are the continuity assumptions. That is, the expected potential outcomes change smoothly as a function of the running variable through the cutoff. In words, this means that nothing that determines the potential outcomes changes abruptly at c_0 except for the treatment assignment. But, this can be violated in practice if:

1. the assignment rule is known in advance
2. agents are interested in adjusting
3. agents have time to adjust

Examples include re-taking an exam, self-reported income, etc. But some other unobservable characteristic change could happen at the threshold, and this has a direct effect on the outcome. In other words, the cutoff is endogenous. An example would be age thresholds used for policy, such as when a person turns 18 and faces more severe penalties for crime. This age threshold both with the treatment (i.e., higher penalties for crime), but is also correlated with variables that affect the outcomes such as graduating from high school, voting rights, etc.

Because of these challenges to identification, a lot of work by econometricians and applied microeconomists has gone to trying

to figure out solutions to these problems. The most influential is a density test by Justin McCrary, now called the McCrary density test [McCrary, 2008]. The McCrary density is used to check for whether units are sorting on the running variable. Imagine that there were two rooms – room A will receive some treatment, and room B will receive nothing. There are natural incentives for the people in room B to get into room A. But, importantly, if they were successful, then the two rooms would look different. Room A would have more observations than room B – thus evidence for the manipulation.

Manipulation on the sorting variable always has that flavor. Assuming a continuous distribution of units, manipulation would mean that more units are showing up just on the other side of the cut off. Formally, if we assume a desirable treatment D and as assignment rule $X \geq c_0$. If individuals sort into D by choosing X such that $X \geq c_0$, then we say individuals are sorting on the running variable.

The kind of test needed to investigate whether manipulation is occurring is a test that checks whether there is bunching of units at the cutoff. In other words, we need a *density* test. McCrary [2008] suggests a formal test where under the null, the density should be continuous at the cutoff point. Under the alternative hypothesis, the density should increase at the kink.⁹⁹ Mechanically, partition the assignment variable into bins and calculate frequencies (i.e., the number of observations) in each bin. Treat the frequency counts as the dependent variable in a local linear regression. If you can estimate the conditional expectations, then you have the data on the running variable, so in principle you can always do a density test. You can download the (no longer supported) Stata ado package DCdensity¹⁰⁰ or the package rddensity, or you can install it for R as well.¹⁰¹

For RDD to be useful, you already need to know something about the mechanism generating the assignment variable and how susceptible it could be to manipulation. Note the rationality of economic actors that this test is built on. A discontinuity in the density is considered suspicious and suggestive of manipulation around the cutoff. This is a high-powered test. You need a lot of observations at c_0 to distinguish a discontinuity in the density from noise. McCrary [2008] presents a helpful picture of a situation with and without manipulation in Figure 47.

There are also helpful visualization of manipulation from other contexts, such as marathon running. Allen et al. [2013] shows a picture of the kinds of density jumps that occur in finishing times. The reason for these finishing time jumps is because many marathon runners have target times that they're shooting for. These are usually 30 minute intervals, but also include unique race qualification

⁹⁹ In those situations, anyway, where the treatment is desirable to the units.

¹⁰⁰ <http://eml.berkeley.edu/~jmcrary/DCdensity/>

¹⁰¹ <http://cran.r-project.org/web/packages/rdd/rdd.pdf>

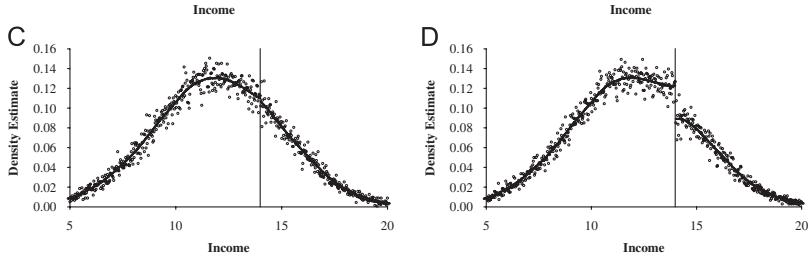


Figure 47: Panel C is density of income when there is no pre-announcement and no manipulation. Panel D is the density of income when there is pre-announcement and manipulation. From McCrary [2008].

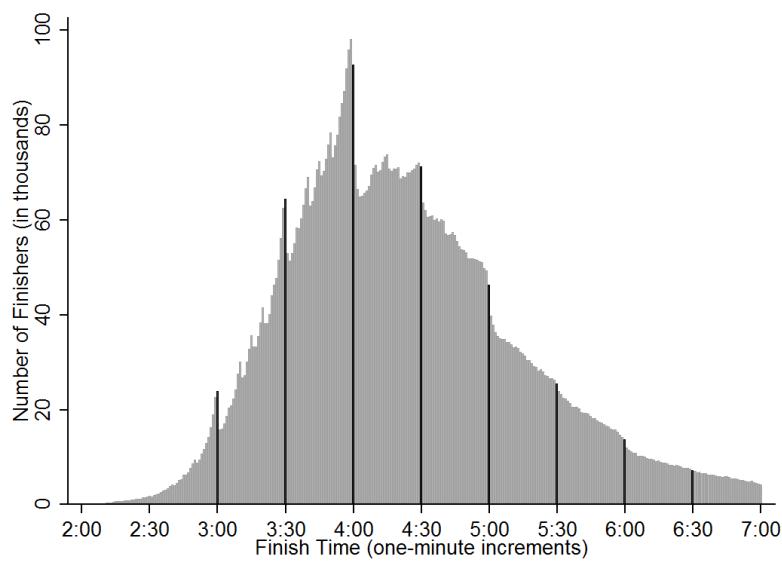
times (e.g., Boston qualifying times). The panel on the top shows a histogram of times with black lines showing jumps in the number of observations. Density tests are provided on the bottom.

Testing for validity It is become common in this literature to provide evidence for the credibility of the underlying identifying assumptions. While the assumptions cannot be directly tested, indirect evidence may be persuasive. We're already mentioned one such test – the McCrary density test. A second test is a covariate balance test. For RDD to be valid in your study, there must not be an observable discontinuous change in the average values of the covariates around the cutoff. As these are pretreatment characteristics, they should be invariant to change in treatment assignment. An example of this is from ¹⁰² where they evaluated the impact of Democratic voteshare, just at 50%, on various demographic factors (Figure 48).

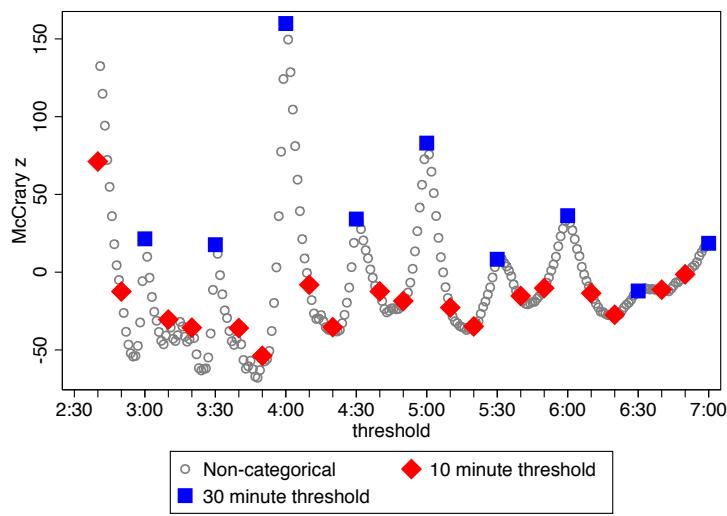
This test is basically what is sometimes called a *placebo* test. That is, you are looking for there to be no effects where there shouldn't be any. So a third kind of test is an extension of that – just as there shouldn't be effects at the cutoff on pretreatment values, there shouldn't be effects on the outcome of interest at arbitrarily chosen cutoffs. Imbens and Lemieux [2008] suggest to look at one side of the discontinuity, take the median value of the running variable in that section, and pretend it was a discontinuity, c'_0 . Then test whether there is a discontinuity in the outcome at c'_0 . You do *not* want to find anything.

¹⁰² David S. Lee. Randomized experiments from non-random selection in u.s. house elections. *Journal of Econometrics*, 142:675–697, 2008

Data visualization RDD papers are intensive data visualization studies. You typically see a lot of pictures. The following are modal. First, a graph showing the outcome variable by running variable is standard. You should construct bins and average the outcome within bins on both sides of the cutoff. You should also look at different bin sizes when constructing these graphs [Lee and Lemieux, 2010]. Plot the running variables X_i on the horizontal axis and the average for Y_i for each bin on the vertical axis. Inspect whether there is a

Figure 2: Distribution of marathon finishing times ($n = 9,378,546$)

NOTE: The dark bars highlight the density in the minute bin just prior to each 30 minute threshold.

Figure 3: Running McCrary z -statistic

NOTE: The McCrary test is run at each minute threshold from 2:40 to 7:00 to test whether there is a significant discontinuity in the density function at that threshold.

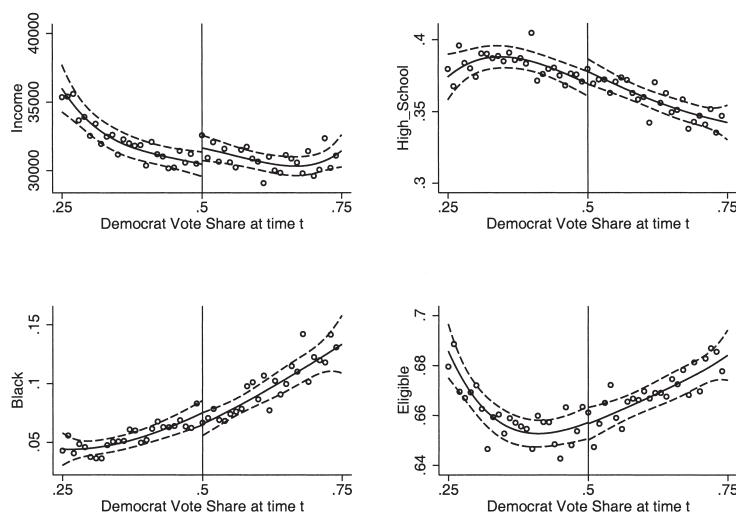


Figure 48: Panels refer to (top left to bottom right) district characteristics: real income, percent high school degree, percent black, and percent eligible to vote. Circles represent the average characteristic within intervals of 0.01 in Democratic vote share. The continuous line represents the predicted values from a fourth-order polynomial in vote share fitted separately for points above and below the 50 percent threshold. The dotted line represents the 95 percent confidence interval.

discontinuity at c_0 . Also inspect whether there are other unexpected discontinuities at other points on X_i . An example of what you want to see is in Figure 49.

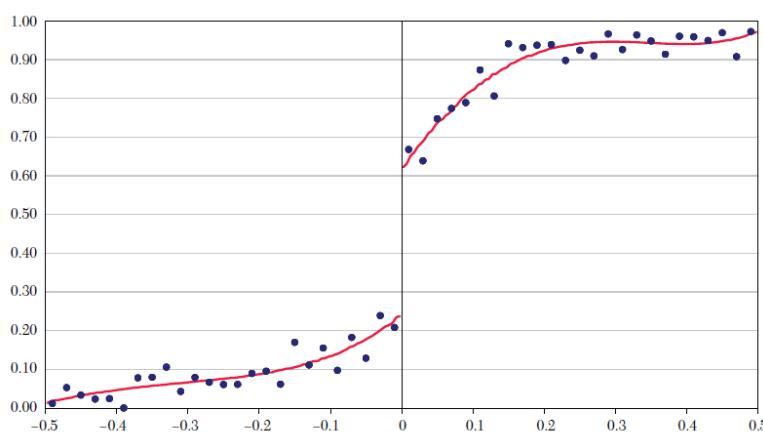


Figure 49: Example of outcome plotted against the running variable.

If it's a fuzzy design, then you want to see a graph showing the probability of treatment jumps at c_0 . This tells you whether you have a first stage. You also want to see evidence from your placebo tests. As I said earlier, there should be no jump in some covariate at c_0 , so readers should be shown this lack of an effect visually, as well as in a regression.

Another graph that is absolutely mandatory is the McCrary density test. The reader must be shown that there is no sign of manip-

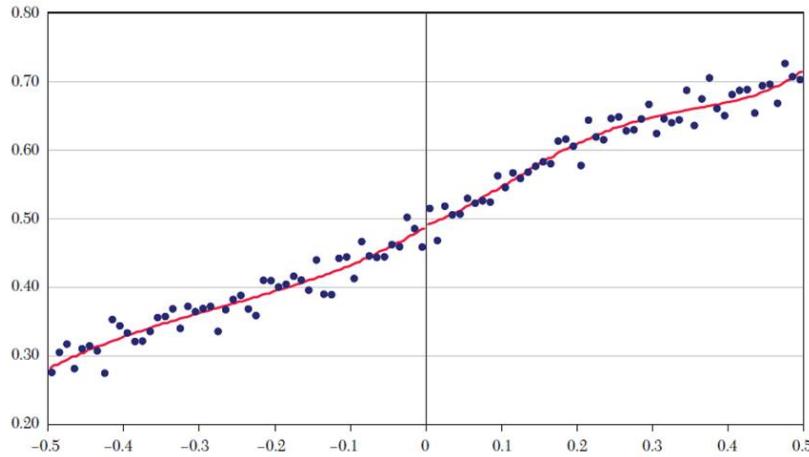


Figure 17. Discontinuity in Baseline Covariate (Share of Vote in Prior Election)

Figure 50: Example of covariate plotted against the running variable.

ulation. One can either use a canned routine to do this, such as `rddensity` or `DCDensity`, or do it oneself. If one does it oneself, then the method is to plot the number of observations into bins. This plots allows us to investigate whether there is a discontinuity in the distribution of the running variable at the threshold. If so, this suggests that people are manipulating the running variable around the threshold. This is an indirect test of the identifying assumption that each individual has imprecise control over the assignment variable. An example of a dataset where manipulation seems likely is the National Health Interview Survey where respondents were asked about participation in the Supplemental Nutrition Assistance Program (SNAP). I merged into the main survey data imputed income data. As SNAP eligibility is based in part on gross monthly income and family size, I created a running variable based on these two variables. Individuals with income that surpassed some given monthly income level appropriate for their family size were then eligible for SNAP. But if there was manipulation, meaning some people misreported their income in order to become eligible for SNAP, we would expect the number of people with income just below that threshold would jump. I estimated a McCrary density test to evaluate whether there was evidence for that. I present that evidence in Figure 51.

That, in fact, is exactly what I find. And statistical tests on this difference are significant at the 1% level, suggesting there is evidence for manipulation.

Example: Elect or Affect [Lee et al., 2004] To illustrate how to implement RDD in practice, we will replicate the Lee et al. [2004] paper.

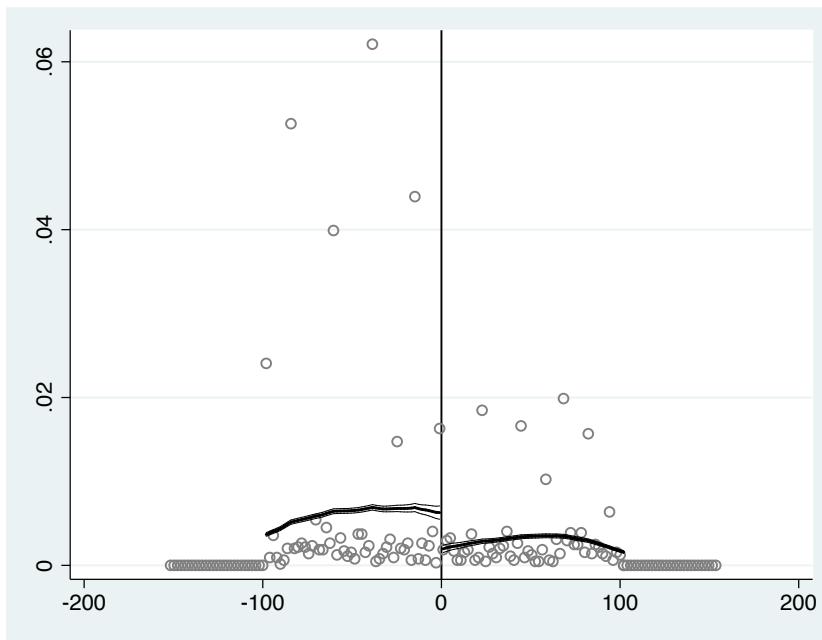


Figure 51: McCrary density test, NHIS data, SNAP eligibility against a running variable based on income and family size.

First install the data. It's large, so it will take a moment to get fully downloaded.

```
. scuse lmb-data
```

The big question motivating this paper has to do with whether and in what way voters affect policy. There are two fundamentally different views of the role of elections in a representative democracy. They are:

1. **Convergence:** Heterogeneous voter ideology forces each candidate to moderate his or her position (e.g., similar to the median voter theorem).

"Competition for votes can force even the most partisan Republicans and Democrats to moderate their policy choices. In the extreme case, competition may be so strong that it leads to 'full policy convergence': opposing parties are forced to adopt identical policies." [Lee et al., 2004]

2. **Divergence:** When partisan politicians cannot credibly commit to certain policies, then convergence is undermined. The result can be fully policy divergence. Divergence is when the winning candidate, after taking office, simply pursues his most-preferred policy. In this case, voters fail to compel candidates to reach any kind of policy compromise.

The authors present a model, which I've simplified. Let R and D be candidates in a Congressional race. The policy space is a single dimension where D and R 's policy preferences in a period are quadratic loss functions, $u(l)$ and $v(l)$, and l is the policy variable. Each player has some bliss point, which is their most preferred location along the unidimensional policy range. For Democrats, it's $l^* = c(> 0)$ and for Republicans it's $l^* = 0$. Here's what this means.

Ex ante, voters expect the candidate to choose some policy and they expect the candidate to win with probability $P(x^e, y^e)$ where x^e and y^e are the policies chosen by Democrats and Republicans, respectively. When $x^e > y^e$, then $\frac{\partial P}{\partial x^e} > 0$, $\frac{\partial P}{\partial y^e} < 0$.

P^* represents the underlying popularity of the Democratic party, or put differently, the probability that D would win if the policy chosen x equalled the Democrat's bliss point c .

The solution to this game has multiple Nash equilibria, which I discuss now.

1. Partial/Complete Convergence: Voters affect policies.

- The key result under this equilibrium is $\frac{\partial x^*}{\partial P^*} > 0$.
- Interpretation: if we dropped more Democrats into the district from a helicopter, it would exogenously increase P^* and this would result in candidates changing their policy positions, i.e., $\frac{\partial x^*}{\partial P^*} > 0$

2. Complete divergence: Voters elect politicians with fixed policies who do whatever they want to do.¹⁰³

- Key result is that more popularity has no effect on policies. That is $\frac{\partial x^*}{\partial P^*} = 0$.
- An exogenous shock to P^* (i.e., dropping Democrats into the district) does *nothing* to equilibrium policies. Voters elect politicians who then do whatever they want because of their fixed policy preferences.

¹⁰³ The "honey badger" don't care. It takes what it wants. See <https://www.youtube.com/watch?v=4r7wHMg5Yjg>.

Potential roll-call voting record outcomes of the representative following some election is

$$RC_t = D_t x_t + (1 - D_t) y_t$$

where D_t indicates whether a Democrat won the election. That is, only the winning candidate's policy is observed. This expression can be transformed into regression equations:

$$\begin{aligned} RC_t &= \alpha_0 + \pi_0 P_t^* + \pi_1 D_t + \varepsilon_t \\ RC_{t+1} &= \beta_0 + \pi_0 P_{t+1}^* + \pi_1 D_{t+1} + \varepsilon_{t+1} \end{aligned}$$

where α_0 and β_0 are constants.

This equation can't be directly estimated because we never observe P^* . But suppose we could randomize D_t . Then D_t would be independent of P_t^* and ε_t . Then taking conditional expectations with respect to D_t we get:

$$\underbrace{E[RC_{t+1}|D_t = 1] - E[RC_{t+1}|D_t = 0]}_{\text{Observable}} = \pi_0[P_{t+1}^{*D} - P_{t+1}^{*R}] + \underbrace{\pi_1[P_{t+1}^D - P_{t+1}^R]}_{\text{Observable}} = \underbrace{\gamma}_{\text{Total effect of initial win on future roll call votes}} \quad (80)$$

$$\underbrace{E[RC_t|D_t = 1] - E[RC_t|D_t = 0]}_{\text{Observable}} = \pi_1 \quad (81)$$

$$\underbrace{E[D_{t+1}|D_t = 1] - E[D_{t+1}|D_t = 0]}_{\text{Observable}} = P_{t+1}^D - P_{t+1}^R \quad (82)$$

The “elect” component is $\pi_1[P_{t+1}^D - P_{t+1}^R]$ and it's estimated as the difference in mean voting records between the parties at time t .

The fraction of districts won by Democrats in $t + 1$ is an estimate of $[P_{t+1}^D - P_{t+1}^R]$. Because we can estimate the total effect, γ , of a Democrat victory in t on RC_{t+1} , we can net out the elect component to implicitly get the “affect” component.

But random assignment of D_t is crucial. For without it, this equation would reflect π_1 and selection (i.e., Democratic districts have more liberal bliss points). So the authors aim to randomize D_t using a RDD, which I'll now discuss in detail.

There are two main datasets in this project. The first is a measure of how liberal an official voted. This is collected from the Americans for Democratic Action (ADA) linked with House of Representatives election results for 1946-1995. Authors use the ADA score for all US House Representatives from 1946 to 1995 as their voting record index. For each Congress, the ADA chose about 25 high-profile roll-call votes and created an index varying from 0 to 100 for each Representative. Higher scores correspond to a more “liberal” voting record.

The running variable in this study is the voteshare. That is the share of all votes that went to a Democrat. ADA scores are then linked to election returns data during that period.

Recall that we need randomization of D_t . The authors have a clever solution. They will use arguably exogenous variation in Democratic wins to check whether convergence or divergence is correct. Their exogenous shock comes from the discontinuity in the running variable. At a voteshare of just above 0.5, the Democratic candidate wins. They argue that just around that cutoff, random chance de-

terminated the Democratic win - hence the random assignment of D_t .

TABLE I
RESULTS BASED ON ADA SCORES—CLOSE ELECTIONS SAMPLE

Variable	Total effect			Elect component	Affect component
	γ	π_1	$(P_{t+1}^D - P_{t+1}^R)$	$\pi_1[(P_{t+1}^D - P_{t+1}^R)]$	$\pi_0[P_{t+1}^{*D} - P_{t+1}^{*R}]$
	ADA_{t+1}	ADA_t	DEM_{t+1}	(col. 2)*(col. 3))	(col. 1)) - (col. 4))
	(1)	(2)	(3)	(4)	(5)
Estimated gap	21.2 (1.9)	47.6 (1.3)	0.48 (0.02)		
				22.84 (2.2)	-1.64 (2.0)

Standard errors are in parentheses. The unit of observation is a district-congressional session. The sample includes only observations where the Democrat vote share at time t is strictly between 48 percent and 52 percent. The estimated gap is the difference in the average of the relevant variable for observations for which the Democrat vote share at time t is strictly between 50 percent and 52 percent and observations for which the Democrat vote share at time t is strictly between 48 percent and 50 percent. Time t and $t + 1$ refer to congressional sessions. ADA_t is the adjusted ADA voting score. Higher ADA scores correspond to more liberal roll-call voting records. Sample size is 915.

Figure 52: Lee, Moretti and Butler (2004)'s Table 1. Main results.

You should have the data in memory, but if not, recall that the command is:

```
. scuse lmb-data
```

First we will replicate the first column of Figure 52 by typing (with output below each command):

```
. reg score lagdemocrat if lagdemvoteshare>.48 & lagdemvote-share<.52, cluster(id)
```

Number of obs = 915

<snip>

score | Coef. Std. Err.

lagdemocrat | 21.28387 1.951234

```
. reg score democrat if lagdemvoteshare>.48 & lagdemvote-share<.52, cluster(id)
```

Number of obs = 915

<snip>

score | Coef. Std. Err.

democrat | 47.7056 1.356011

```
. reg democrat lagdemocrat if lagdemvoteshare>.48 & lagdemvote-share<.52, cluster(id)
```

Number of obs = 915

<snip>

democrat | Coef. Std. Err.

```
-----+  
lagdemocrat | .4843287 .0289322
```

Okay, a few things. First, notice the similarity between each regression output and the regression output in Figure 52. So as you can see, when we say we are estimating global regressions, it means we are simply regressing some outcome onto a treatment variable. Here what we did was simply run “local” linear regressions, though. Notice the bandwidth - we are only using observations between 0.48 and 0.52 voteshare. So this regression is estimating the coefficient on D_t right around the cutoff. What happens if we use all the data?

```
. reg score democrat, cluster(id2)
```

Number of obs = 13588

score | Coef. Std. Err.

```
-----+  
democrat | 40.76266 1.495659
```

Notice when we use all the data, the effect on the democrat variable becomes smaller. It remains significant, but it no longer includes in its confidence interval the coefficient we found earlier.

Recall we said that it is common to center the running variable. Centering simply means subtracting from the running variable the value of the cutoff so that values of 0 are where the voteshare equals 0.5, negative values are Democratic voteshares less than 0.5, and positive values are Democratic voteshares above 0.5. To do this, type in the following lines:

```
. gen demvoteshare_c = demvoteshare - 0.5  
. reg score democrat demvoteshare_c, cluster(id2)
```

Number of obs = 13577

score | Coef. Std. Err.

democrat		58.50236	1.555847
demvoteshare_c		-48.93761	4.441693

Notice, now controlling for the running variable causes the coefficient on democrat – using all the data – to get much larger than when we didn't control for the running variable.

It is common, though, to allow the running variable to vary on either side of the discontinuity, but how exactly do we implement that? Think of it - we need for a regression line to be on either side, which means necessarily that we have *two* lines left and right of the discontinuity. To do this, we need an interaction - specifically an interaction of the running variable with the treatment variable. So to do that in Stata, we simply type:

```
. xi: reg score i.democrat*demvoteshare_c, cluster(id2)
```

Number of obs = 13577

<snip>

score | Coef. Std. Err.

_Idemocrat_1		55.43136	1.448568
demvoteshare_c		-5.682785	5.939863
_IdemXdemvo_1		-55.15188	8.236231

But notice, we are still estimating *global* regressions. And it is for that reason, as I'll show now, that the coefficient is larger. This suggests that there exist strong outliers in the data which are causing the distance at c_0 to spread more widely. So a natural solution, therefore, is to again limit our analysis to a smaller window. What this does is drop the observations far away from c_0 , and therefore omit the influence of outliers from our estimation at the cutoff. Since we used $+/- .02$ last time, we'll use $+/- .05$ this time just to mix things up.

```
. xi: reg score i.democrat*demvoteshare_c if demvoteshare>.45  
& demvoteshare<.55, cluster(id2)
```

Number of obs = 2387

<snip>

score | Coef. Std. Err.

```
_Idemocrat_1 | 46.77845 2.491464
demvoteshare_c | 54.82604 50.12314
_IdemXdemvo_1 | -91.1152 81.05893
```

As can be seen, when we limit our analysis to $+/- 0.05$ around the cutoff, we are dropping observations from the analysis. That's why we only have 2,387 observations for analysis as opposed to the 13,000 we had before. This brings us to an important point. The ability to do this kind of local regression analysis necessarily requires a lot of data *around* the cutoff. If we don't have a lot of data around the cutoff, then we simply cannot estimate local regression models, as the data simply becomes too noisy. This is why I said RDD is "greedy". It needs a lot of data because it uses only a portion of it for analysis.

But putting that aside, think about what this did. This fit a model where it controlled for a straight line below the cutoff (demvote-share_c) and above the cutoff (_IdemXdemvo_1). Controlling for those two things, the remainder is a potential gap at voteshare=0.5, which is captured by the democrat dummy. It does this through extrapolation.

I encourage you to play around with the windows. Try $+/- 0.1$ and $+/- 0.01$. Notice how the standard errors get larger the more narrow you make the band. Why do you think that is? Think about that - narrowing the band decreases bias, but strangely increases variance. Do you know why?

Recall what we said about nonlinearities and strong trends in the evolution of the potential outcomes. Without controlling for nonlinearities, we may be misattributing causal effects using only linear functions of the running variable. Therefore next we will show how to model polynomials in the running variable. [Gelman and Imbens \[2017\]](#) recommend polynomials up to a quadratic to avoid the problem of overfitting. So we will follow their advice now.

First, we need to generate the polynomials. Then we need to interact them with the treatment variable which as we alluded to earlier will allow us to model polynomials to the left and right of the cutoff.

```
. gen x_c = demvoteshare - 0.5
. gen x_c2 = x_c^2
. reg score democrat##(c.x_c c.x2_c)
```

Number of obs = 13,577

score | Coef. Std. Err.

```

1.democrat | 44.40229 1.008569
x_c | -23.8496 8.209109
x_c2 | -41.72917 17.50259
democrat#c.x_c |
1 | 111.8963 10.57201
democrat#c.x_c2 |
1 | -229.9544 21.10866

```

Notice now that using all the data gets us closer to the estimate.
And finally, we can use the a narrow bandwidth.

```

. reg score democrat##(c.x_c c.x2_c) if demvoteshare>0.4
& demvoteshare<0.6
Number of obs = 4,632

score | Coef. Std. Err.
-----+
1.democrat | 45.9283 1.892566
x_c | 38.63988 60.77525
x_c2 | 295.1723 594.3159
democrat#c.x_c |
1 | 6.507415 88.51418
democrat#c.x_c2 |
1 | -744.0247 862.0435

```

Once we controlled for the quadratic polynomial, the advantage of limiting the bandwidth was a smaller than when we were either not controlling for the running variable at all or controlling only a linear running variable.

Hahn et al. [2001] clarified assumptions about RDD – specifically, that continuity of the conditional expected potential outcomes. They also framed estimation as a non-parametric problem and emphasized using local polynomial regressions.

Nonparametric methods mean a lot of different things to different people in statistics, but in RDD contexts, the idea is to estimate a model that doesn't assume a functional form for the relationship between the outcome variable (Y) and the running variable (X). The model would be something like this:

$$Y = f(X) + \varepsilon$$

A very basic method would be to calculate $E[Y]$ for each bin on X , like a histogram. And Stata has an option to do this called cmogram

created by Christopher Robert. The program has a lot of useful options, and we can recreate Figures I, IIA and IIB from Lee et al. [2004]. Here is Figure I which is the relationship between the democratic win (as a function of the running variable, democratic vote share) and the candidates second period ADA score (Figure 53).

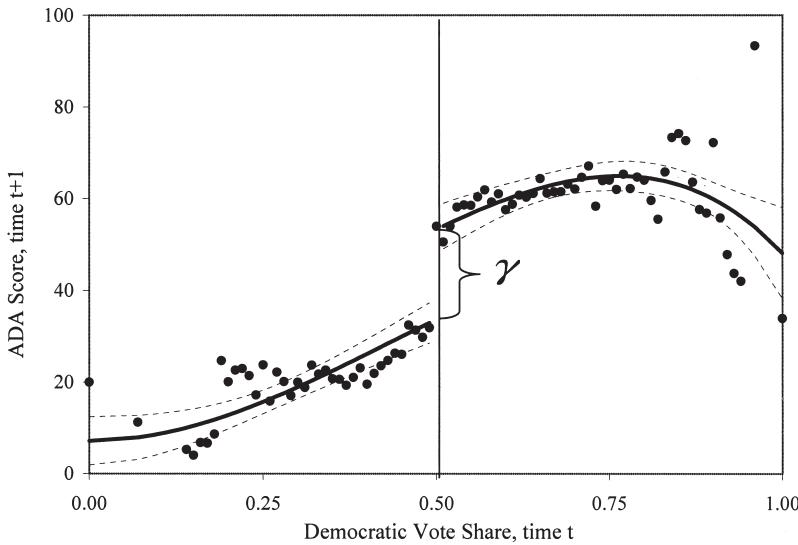


Figure 53: Lee et al. [2004], Figure I

FIGURE I

Total Effect of Initial Win on Future ADA Scores: γ

This figure plots ADA scores after the election at time $t + 1$ against the Democrat vote share, time t . Each circle is the average ADA score within 0.01 intervals of the Democrat vote share. Solid lines are fitted values from fourth-order polynomial regressions on either side of the discontinuity. Dotted lines are pointwise 95 percent confidence intervals. The discontinuity gap estimates

$$\gamma = \underbrace{\pi_0(P_{t+1}^{*D} - P_{t+1}^{*R})}_{\text{"Affect"}} + \underbrace{\pi_1(P_{t+1}^{*D} - P_{t+1}^{*R})}_{\text{"Elect"}}$$

First you will need to install `cmogram` from `ssc`, the Statistical Software Components archive.

```
. ssc install cmogram
```

Next we calculate the conditional mean values for the observations according to an automated binning algorithm generated by `cmogram`.

```
. cmogram score lagdemvoteshare, cut(0.5) scatter line(0.5)
qfitci
```

Figure 54 shows the output from this program. Notice the similarities between what we produced here and what Lee et al. [2004] produced in their Figure I. The only difference is subtle differences in the binning used for the two figures. The key arguments used in this

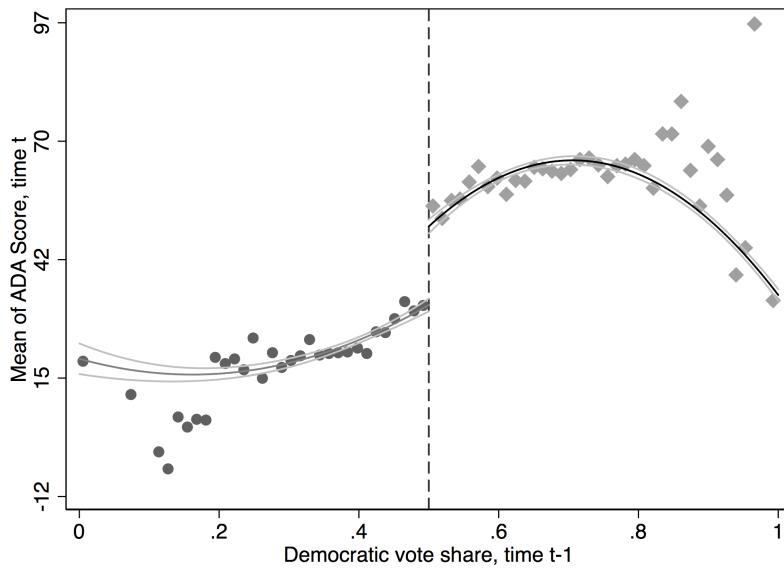


Figure 54: Reproduction of Lee et al. [2004] Figure I using `cmogram` with quadratic fit and confidence intervals

command are the listing of the outcome (`score`) and the running variable (`lagdemvoteshare`), the designation of where along the running variable the cutoff is (`cut(0.5)`), whether to produce the visualization of the scatter plots (`scatter`), whether to show a dashed vertical line at the cutoff (`line(0.5)`) and what kind of polynomial to fit left and right of the cutoff (`qfitci`).

We have options other than a quadratic fit, though, and I think it's useful to compare this graph with one where we only fit a linear model. Now because there are strong trends in the running variable, we probably just want to use the quadratic, but let's see what we get when we use simple lines.

```
. cmogram score lagdemvoteshare, cut(0.5) scatter line(0.5)
lfit
```

Figure 55 shows what we get when we only use a linear fit of the data left and right of the cutoff. Notice the influence that outliers far from the actual cutoff play in the estimate of the causal effect at the cutoff. Now some of this would go away if we restricted the bandwidth to be shorter distances to and from the cutoff, but I leave it to you to do that yourself.

Finally, we can use a lowess fit. A lowess fit more or less crawls through the data running small regressions on small cuts of data. This can give the picture a zig zag appearance. We nonetheless show it here:

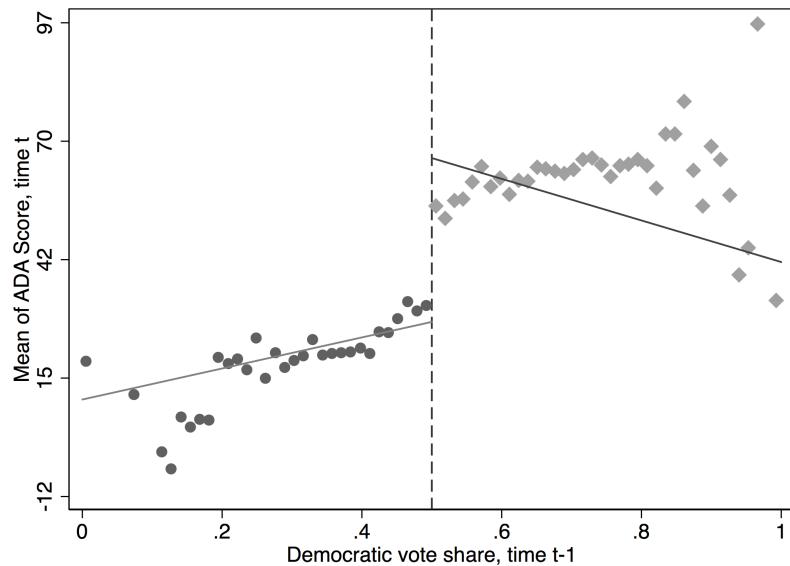


Figure 55: Reproduction of Lee et al. [2004] Figure I using cmogram with linear fit

```
. cmogram score lagdemvoteshare, cut(0.5) scatter line(0.5)
lowess
```

It is probably a good idea to at least run all of these, but your final selection of what to report as your main results should be that polynomial that best fits the data. Some papers only report a linear fit because there weren't very strong trends to begin with. For instance, consider Carrell et al. [2011]. The authors are interested in the causal effect of drinking on academic test outcomes. Their running variable is the precise age of the student, which they have because they know the student's date of birth and they know the date of every exam taken at the Air Force Academy. Because the Air Force Academy restricts the social lives of its students, there is a more stark increase in drinking at age 21 on its campus than might be on a typical university campus. They examined the causal effect of drinking age on normalized grades using RDD, but because there weren't strong trends in the data, they only fit a linear model (Figure 57).

It would no doubt have been useful for this graph to include confidence intervals, but the authors did not. Instead, they estimated

As can be seen from both the graphical data and the regression analysis, there appears to be a break in the outcome (normalized

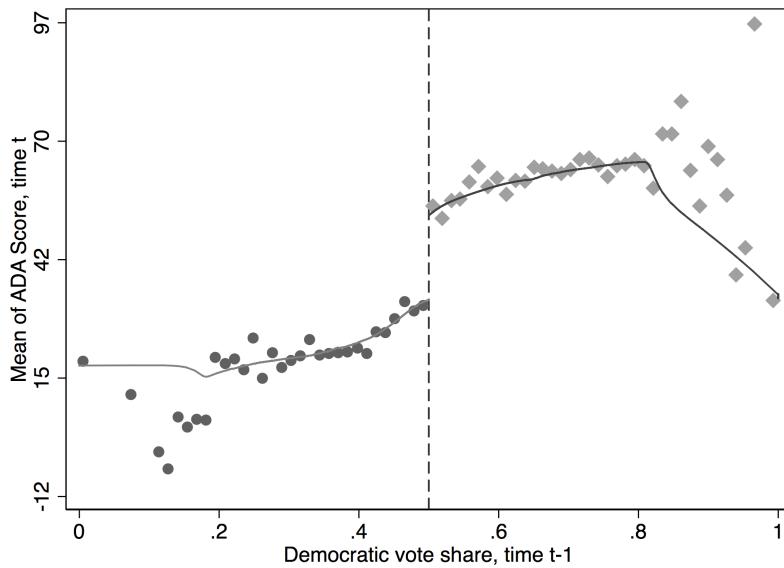


Figure 56: Reproduction of Lee et al. [2004] Figure I using cmogram with lowess fit

grade) at the point of age 21, suggesting that alcohol has a negative causal effect on academic performance.¹⁰⁴

Hahn et al. [2001] have shown that the one-sided kernel estimation estimation such as lowess may suffer from poor properties because the point of interest is at the boundary (i.e., the discontinuity). This is called the “boundary problem”. They propose instead to use “local linear nonparametric regressions” instead. In these regressions, more weight is given to the observations at the center.

You can implement this using Stata’s poly command which estimates kernel-weighted local polynomial regressions. Think of it as a weighted regression restricted to a window like we’ve been doing (hence the word “local”) where the chosen kernel provides the weights. A rectangular kernel would give the same results as $E[Y]$ at a given bin on X , but a triangular kernel would give more importance to observations closest to the center. This method will be sensitive to how large the bandwidth, or window, you choose. But in that sense, it’s similar to what we’ve been doing.

```
. * Note kernel-weighted local polynomial regression is a
smoothing method.
. lpoly score demvoteshare if democrat == 0, nograph
kernel(triangle) gen(x0 sdem0) bwidth(0.1)
. lpoly score demvoteshare if democrat == 1, nograph
kernel(triangle) gen(x1 sdem1) bwidth(0.1)
. scatter sdem1 x1, color(red) msize(small) || scatter
```

¹⁰⁴ Many, many papers have used RDD to look at alcohol using both age as the running variable or blood alcohol content as the running variable. Examples include Carpenter and Dobkin [2009] and Hansen [2015] just to name a couple.

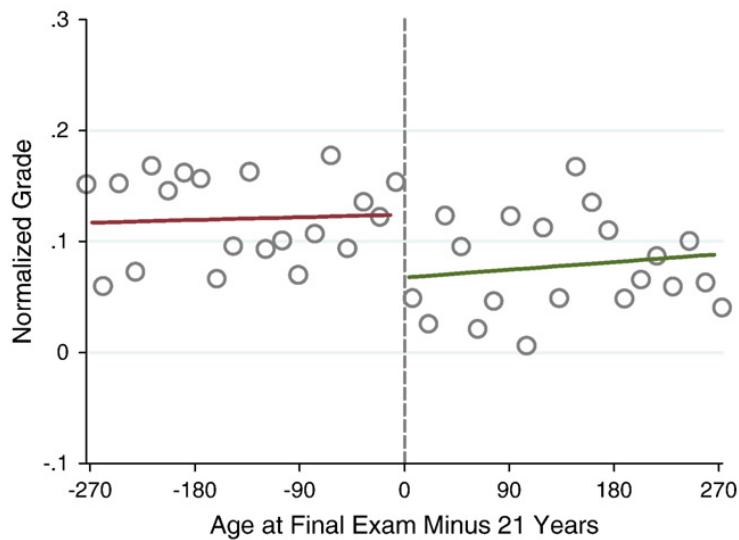


Figure 57: Carrell et al. [2011] Figure 3

Table 3
Regression discontinuity estimates of the effect of drinking on academic performance.

Specification	1	2	3
Discontinuity at age 21	-0.092 *** (0.03)	-0.114 *** (0.02)	-0.106 *** (0.03)
Observations	38,782	38,782	38,782
Age polynomial	Linear	Linear	Quadratic
Control variables	No	Yes	Yes

Figure 58: Carrell et al. [2011] Table 3

```
sdem0 x0, msize(small) color(red) xline(0.5,lstyle(dot))
legend(off) xtitle("Democratic vote share") ytitle("ADA score")
```

Figure 60 shows this visually.

A couple of final things. First, I'm not showing this, but recall the continuity assumption. Because the continuity assumption specifically involves continuous conditional expectation functions of the potential outcomes throughout the cutoff, it therefore is *untestable*. That's right – it's an untestable assumption. But, what we can do is check for whether there are changes in the conditional expectation functions for other exogenous covariates that cannot or should not be changing as a result of the cutoff. So it's very common to look at things like race or gender around the cutoff. You can use these same methods to do that, but I do not do them here. Any RDD paper will

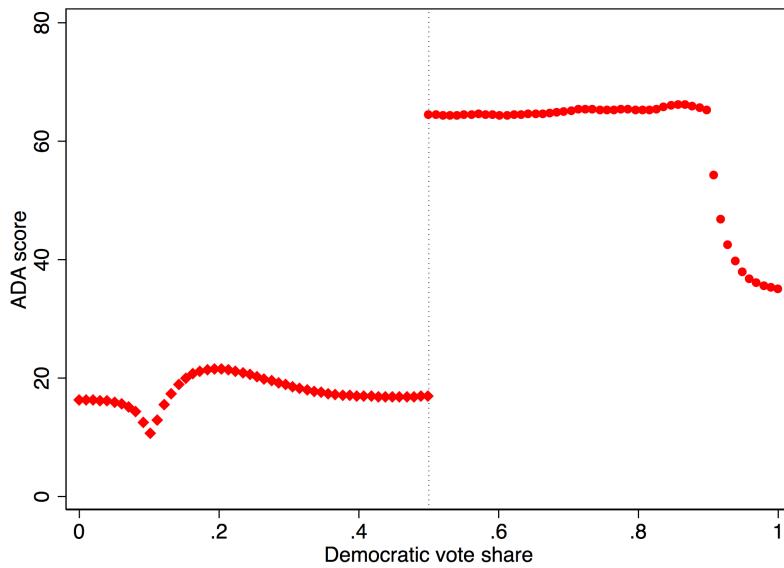


Figure 59: Local linear nonparametric regressions

always involve such placebos; even though they are not direct tests of the continuity assumption, they are indirect tests. Remember, your reader isn't as familiar with this thing you're studying, so your task is teach them. Anticipate their objections and the sources of their skepticism. Think like them. Try to put yourself in a stranger's shoes. And then test those skepticisms to the best of your ability.

Second, we saw the importance of bandwidth selection, or window, for estimating the causal effect using this method, as well as the importance of selection of polynomial length. There's always a tradeoff when choosing the bandwidth between bias and variance - the shorter the window, the lower bias, but because you have less data, the variance in your estimate increases. Recent work has been focused on optimal bandwidth selection, such as [Imbens and Kalyanaraman \[2011\]](#) and [Calonico et al. \[2014\]](#). The latter can be implemented with the user-created `rdrobust` command. These methods ultimately choose optimal bandwidths which may differ left and right of the cutoff based on some bias-variance tradeoff. Here's an example:

```
. ssc install rdrobust
. rdrobust score demvoteshare, c(0.5)
```

Sharp RD estimates using local polynomial regression.
Cutoff c = .5 | Left of c Right of c Number of obs = 13577
+----- BW type = mserd
Number of obs | 5480 8097 Kernel = Triangular

Eff. Number of obs | 2096 1882 VCE method = NN

Order est. (p) | 1 1

Order bias (q) | 2 2

BW est. (h) | 0.085 0.085

BW bias (b) | 0.140 0.140

rho (h/b) | 0.607 0.607

Outcome: score. Running variable: demvoteshare.

Method | Coef. Std. Err.

Conventional | 46.483 1.2445

Robust | - - 31.3500

This method, as we've repeatedly said, is data greedy because it gobbles up data at the discontinuity. So ideally these kinds of methods will be used when you have large numbers of observations in the sample so that you have a sizable number of observations at the discontinuity. When that is the case, there should be some harmony in your findings across results. If there isn't, then it calls into question whether you have sufficient power to pick up this effect.

Finally, we look at the implementation of the McCrary density test. Justin McCrary has graciously made this available to us, though `rdrobust` also has a density test built into it. But for now, we will use McCrary's ado package. This cannot be downloaded from `ssc`, so you must download it directly from McCrary's website and move it into your Stata subdirectory that we listed earlier. The website is <https://eml.berkeley.edu/~jmccrary/DCdensity/DCdensity.ado>. Note this will automatically download the file.

Once the file is installed, you use the following command to check for whether there is any evidence for manipulation in the running variable at the cutoff.

```
. DCdensity demvoteshare_c if (demvoteshare_c>-0.5 &
demvoteshare_c<0.5), breakpoint(0) generate(Xj Yj r0 fhat
se_fhat)
```

Using default bin size calculation, bin size = .003047982

Using default bandwidth calculation, bandwidth = .104944836

Discontinuity estimate (log difference in height): .011195629

(.061618519)

Performing LLR smoothing.

296 iterations will be performed

And visually inspecting the graph, we see no signs that there was manipulation in the running variable at the cutoff.

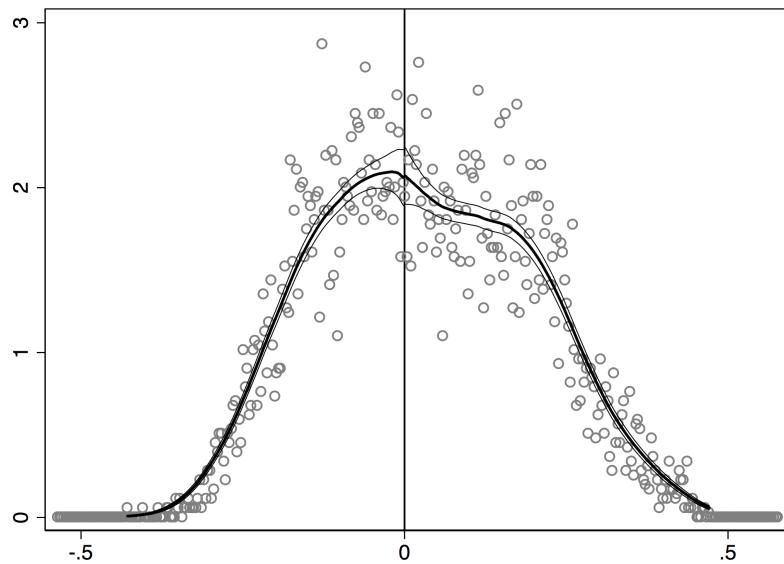


Figure 6o: Local linear nonparametric regressions

Regression Kink Design

A couple of papers came out by David Card and coauthors. The most notable is [Card et al. \[2015\]](#). This paper introduced us to a new method called regression kink design, or RKD. The intuition is rather simple. Rather than the discontinuity creating a discontinuous jump in the treatment variable at the cutoff, it created a change in the first derivative. They use essentially a “kink” in some policy rule to identify the causal effect of the policy using a jump in the first derivative.

Their paper applies the design to answer the question whether the level of unemployment benefits affects the length of time spent unemployed in Austria. Here’s a brief description of the policy. Unemployment benefits are based on income in a base period. The benefit formula for unemployment exhibits two kinks. There is a minimum benefit level that isn’t binding for people with low earnings. Then benefits are 55% of the earnings in the base period. Then there is a maximum benefit level that is adjusted every year. People with dependents get small supplements, which is the reason there are five “solid” lines in the following graph. Not everyone receives benefits that correspond one to one to the formula because mistakes are made in the administrative data (Figure 61).

The graph shows unemployment benefits on the vertical axis as a function of pre-unemployment earnings on the horizontal axis. Next we look at the relationship between average daily unemploy-

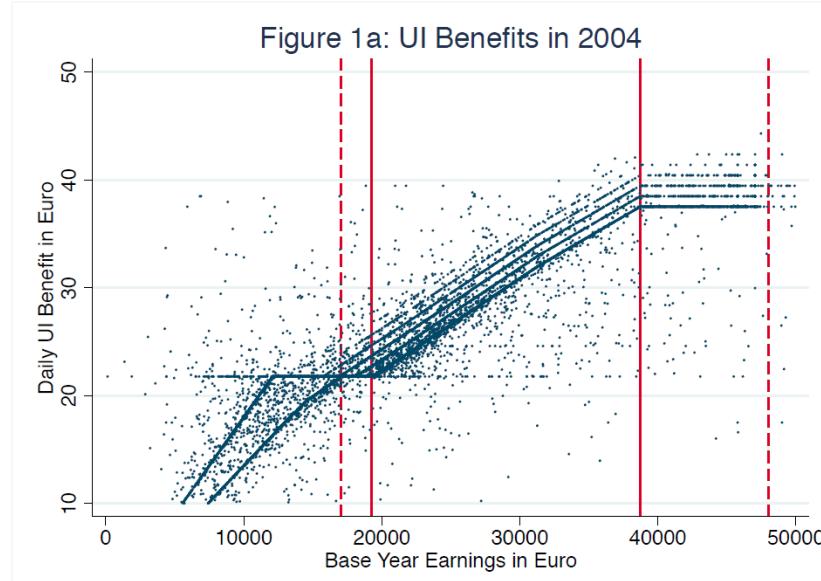


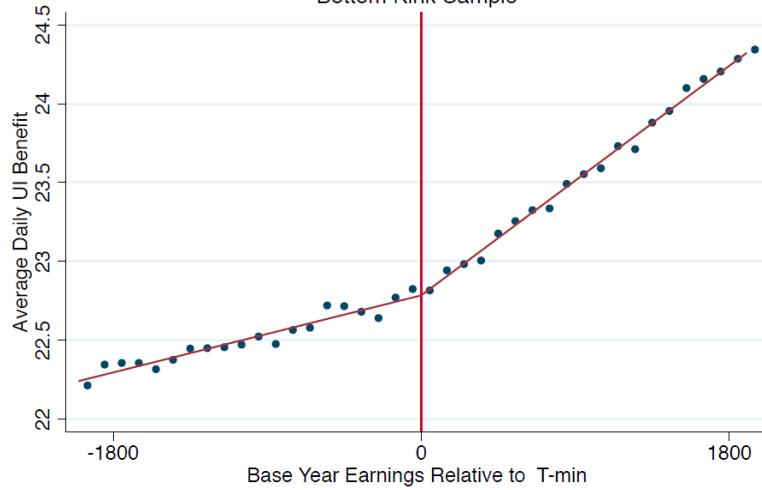
Figure 61: RKD kinks from Card et al. [2015]

ment insurance benefits and base year earnings, where the running variable has re-centered. The bin-size is 100 euros. For single individuals unemployment insurance benefits are flat below the cutoff. The relationship is still upward sloping, though, because of family benefits.

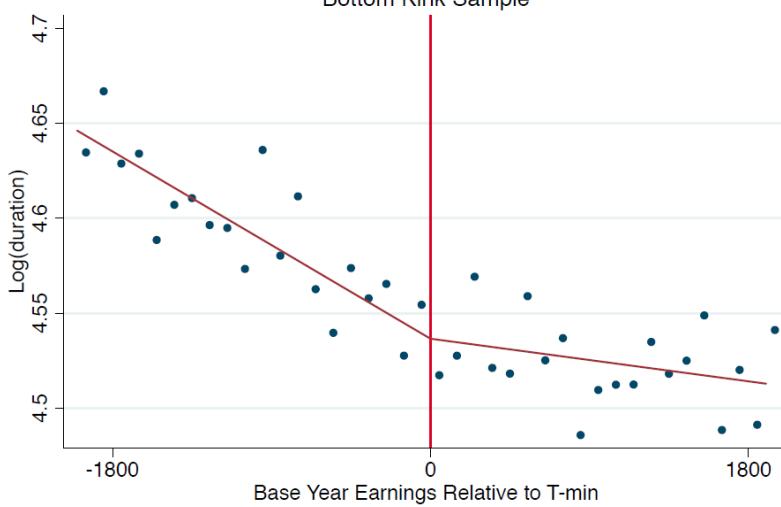
Next we look at the main outcome of interest – time unemployed, which is the time the individual spent until they got another job. As can be seen in Figure 63, people with higher base earnings have less trouble finding a job (which gives it the negative slope). But there is a kink - the relationship becomes shallower once benefits increase more. This suggests that as unemployment benefits increased, the time spent unemployed was longer – even though it continued to rise, the slope shifted and got flatter. A very interesting and policy-relevant result.

Figure 3: Daily UI Benefits

Bottom Kink Sample

**Figure 62: Base year earnings and benefits for single individuals from Card et al. [2015]****Figure 5: Log Time to Next Job**

Bottom Kink Sample

**Figure 63: Log(duration unemployed) and benefits for single individuals from Card et al. [2015]**

Instrumental variables

Instrumental variables is maybe one of most important econometric strategies ever devised. Just as Archimedes said “Give me a fulcrum, and I shall move the world”, so it could be said that with a good enough instrument, we can identify any causal effect.

But, while that is hyperbole for reasons we will soon see, it is nonetheless the case that instrumental variables is an important contribution to causal inference, and an important tool to have in your toolkit. It is also, interestingly, unique because it is one of those instances where the econometric estimator was not simply ripped off from statistics (e.g., Eicker-Huber-White standard errors) or some other field (e.g., regression discontinuity). Its history is, in my opinion, quite fascinating, and before we dive into the technical material, I’d like to tell you a story about its discovery.

History of Instrumental Variables: Father and Son

Philip Wright was born in 1861 and died in 1934. He received his bachelor’s degree from Tufts in 1884 and a masters degree from Harvard in 1887.¹⁰⁵ His son, Sewall Wright, was born in 1889 when Philip was 28. The family moved from Massachusetts to Illinois where Philip took a position as professor of mathematics and economics at Lombard College. Philip was so unbelievably busy with teaching and service that is astonishing he had any time for research, but he did. He published numerous articles and books over his career, including poetry. You can see his vita here at https://scholar.harvard.edu/files/stock/files/wright_cv.pdf.¹⁰⁶ Sewell attended Lombard College and took his college mathematics courses from his father.

In 1913, Philip took a position at Harvard, and Sewell entered as a graduate student. Eventually Philip would leave for the Brookings Institute, and Sewell would take his first job in the Department of Zoology at the University of Chicago where he would eventually be promoted to professor in 1930.

¹⁰⁵ This biographical information is drawn from Stock and Trebbi [2003].

¹⁰⁶ Interesting side note: Philip had a passion for poetry, and even published some in his life, and he used his school’s printing press to publish the first book of poems by the great American poet, Carl Sandburg.

Philip was prolific which given his teaching and service requirements is amazing. He published in top journals such as the *Quarterly Journal of Economics*, *Journal of the American Statistical Association*, *Journal of Political Economy* and *American Economic Review*. A common theme across many publications was the identification problem. He was acutely aware of it and was intent on solving it.

In 1928, Philip was writing a book about animal and vegetable oils of all the things. The reason? He believed that recent tariff increases were harming international relations. Thus he wrote passionately about the damage from the tariffs, which affected animal and vegetable oils. We will return to this book again, as it's an important contribution to our understanding of instrumental variables.

While Philip was publishing like a fiend in economics, Sewall Wright was revolutionizing the field of genetics. He invented path analysis, a precursor to Pearl's directed acyclical graphical models, as well as made important contributions to the theory of evolution and genetics. He was a genius.

The decision to not follow in the family business (economics) created a bit of tension between the two men, but all evidence suggests that they found one another intellectually stimulating. In his book on vegetable and oil tariffs, there is an Appendix (entitled Appendix B) in which the calculus of the instrumental variables estimator was worked out. Elsewhere, Philip thanked his son for his valuable contributions to what he had written, referring to the path analysis primarily which Sewell taught him. This path analysis, it turned out, played a key role in Appendix B.

The Appendix shows a solution to the identification problem. So long as the economist is willing to impose some restrictions on the problem, then the system of equations can be identified. Specifically, if there is one instrument for supply, and the supply and demand errors are uncorrelated, then the elasticity of demand can be identified.

But who wrote this Appendix B? Either man could've done so. It is an economics article, which points to Philip. But it used the path analysis, which points to Sewell. Historians have debated this, even going so far as to accuse Philip of stealing the idea from his son. If Philip stole the idea, by which I mean when he published Appendix B, he failed to give proper attribution to his son, then it would at the very least have been a strange oversight which was possibly out of character for a man who by all evidence loved his son very much. In comes [Stock and Trebbi \[2003\]](#).

[Stock and Trebbi \[2003\]](#) tried to determine the authorship of Appendix B using "stylometric analysis". Stylometric analysis had been used in other applications, such as to identify the author of

the political novel *Primary Colors* (Joseph Klein) and the unsigned *Federalist Papers*. But Stock and Trebbi [2003] is the first application of it in economics to my knowledge.¹⁰⁷

The method is akin to contemporary machine learning methods. The authors collected raw data containing the known original academic writings of each man, plus the first chapter and Appendix B of the book in question. The writings were edited to exclude footnotes, graphs and figures. Blocks of 1,000 words were selected from the files. A total of 54 blocks were selected: 20 written by Sewall with certainty, 25 by Philip, six from Appendix B, and three from chapter 1. Chapter 1 has always been attributed to Philip, but Stock and Trebbi [2003] treat the three blocks as unknown to “train” the data. That is, they use it to check if their model is correctly predicting authorship.

The stylometric indicators that they used included the frequency of occurrence in each block of 70 function words. The list was taken from a separate study. These 70 function words produced 70 numerical variables, each of which is a count, per 1,000 words, of an individual function word in the block. Some words were dropped because they occurred only once (“things”), leaving 69 function word counts.

The second set of stylometric indicators, taken from another study, concerned grammatical constructions. Stock and Trebbi [2003] used 18 grammatical constructions, which were frequency counts. They included things like noun followed by an adverb, total occurrences of prepositions, coordinating conjunction followed by noun, and so on. There was one dependent variable in their analysis, and that was authorship. The independent variables were 87 covariates (69 function word counts and 18 grammatical statistics).

The results of this analysis are absolutely fascinating. For instance, many covariates have very large *t*-statistics, which would be unlikely if there really were no stylistic differences between the authors and indicators were independently distributed.

So what do they find. The results that I find the most interesting is their regression analysis. They write:

“We regressed authorship against an intercept, the first two principal components of the grammatical statistics and the first two principal components of the function word counts, and we attribute authorship depending on whether the predicted value is greater or less than 0.5.”

Note, they used principal component analysis because they had more covariates than observations, and needed the dimension reduction. A more contemporary method might be LASSO or ridge regression. But, given this analysis, what did they find? They found that all of the Appendix B and chapter 1 blocks were assigned to Philip, not

¹⁰⁷ Maybe the only one?

Sewell. They did other robustness checks, and all of them point to Philip as the author.

I love this story for many reasons. First, I love the idea that an econometric estimator as important as instrumental variables was in fact created by an economist. I'm so accustomed to stories in which the actual econometric estimator was lifted from statistics (Huber-White standard errors) or educational psychology (regression discontinuity). It is nice to know economists have added their own to the seminal canon of econometrics. But the other part of the story that I love is the father/son component. I find it encouraging to know that a father and son can overcome differences through intellectual collaborations such as this. Such relationships are important, and tensions, when they arise, should be vigorously pursued until those tensions to dissipate if possible. And Philip and Sewell give a story of that, which I appreciate.

Natural Experiments and the King of the North

While natural experiments are not technically instrumental variables estimator, they can be construed as such if we grant that they are *the reduced form* component of the IV strategy. I will begin by describing one of the most famous, and my favorite, example of a natural experiment - John Snow's discovery that cholera was a water borne disease transmitted through the London water supply.

Natural experiments are technically, though, not an estimator or even an experiment. Rather they are usually nothing more than an event that occurs naturally which causes exogenous variation in some treatment variable of interest.¹⁰⁸

When thinking about these, effort is spent finding some rare circumstance such that a consequential treatment was handed to some people or units but denied to others "haphazardly". Note I did not say randomly, though ideally it was random or conditionally random. Rosenbaum [2010] wrote:

"The word 'natural' has various connotations, but a 'natural experiment' is a 'wild experiment' not a 'wholesome experiment,' natural in the way that a tiger is natural, not in the way that oatmeal is natural."

Before John Snow was the King of the North, he was a 19th century physician in London during several cholera epidemics. He watched helplessly as patient after patient died from this mysterious illness. Cholera came in waves. Tens of thousands of people died horrible deaths from this disease, and doctors were helpless at stopping it, let alone understanding why it was happening. Snow tried his best to save his patients, but despite that best, they still died.

¹⁰⁸ Instruments don't have to be simply naturally occurring random variables. Sometimes they are lotteries, such as in the Oregon Medicaid Experiment. Other times, they are randomized peer designs to induce participation in an experiment.

Best I can tell, Snow was fueled by compassion, frustration and curiosity. He observed the progression of the disease and began forming conjectures. The popular theory of the time was *miasmis*. *Miasmis* was the majority view about disease transmission, and proponents of the theory claimed that minute, inanimate particles in the air were what caused cholera to spread from person to person. Snow tried everything he could to block the poisons from reaching the person's body, a test of *miasmis*, but nothing seemed to save his patients. So he did what any good scientist does - he began forming a new hypothesis.

It's important to note something: cholera came in three waves in London, and Snow was there for all of them. He was on the front line, both as a doctor and an epidemiologist. And while his patients were dying, he was paying attention - making guesses, testing them, and updating his beliefs along the way.

Snow observed the clinical course of the disease and made the following conjecture. He posited that the active agent was a living organism that entered the body, got into the alimentary canal with food or drink, multiplied in the body, and generated a poison that caused the body to expel water. The organism passed out of the body with these evacuations, then entered the water supply, re-infected new victims who unknowingly drank from the water supply. This process repeated causing a cholera epidemic.

Snow had evidence for this based on years of observing the progression of the disease. For instance, cholera transmission tended to follow human commerce. Or the fact that a sailor on a ship from a cholera-free country who arrived at a cholera-stricken port would only get sick after landing or taking on supplies. Finally, cholera hit the poorest communities the worst, who also lived in the most crowded housing with the worst hygiene. He even identified Patient Zero - a sailor named John Harnold who arrived to London by the Elbe ship from Hamburg where the disease was prevailing.

It seems like you can see Snow over time moving towards cleaner and cleaner pieces of evidence in support of a waterborne hypothesis. For instance, we know that he thought it important to compare two apartment buildings - one which was heavily hit with cholera, but a second one that wasn't. The first building was contaminated by runoff from privies but the water supply in the second was cleaner. The first building also seemed to be hit much harder by cholera than the second. These facts, while not entirely consistent with the miasma theory, were still only suggestive.

How could Snow test a hypothesis that cholera was transmitted via poisoned water supplies? Simple! Randomly assign half of London to drink from water contaminated by the runoff from cholera

victims, and the other half from clean water. But it wasn't merely that Snow predated the experimental statisticians, Jerzy Neyman and Roland Fisher, that kept him from running an experiment like that. An even bigger constraint was that even if he had known about randomization, there's no way he could've run an experiment like that. Oftentimes, particularly in social sciences like epidemiology and economics, we are dealing with macro-level phenomena and randomized experiments are simply not realistic options.

I present that kind of thought experiment, though, not to advocate for the randomized controlled trial, but rather to help us understand the constraints we face, as well as to help hone in on what sort of experiment we need in order to test a particular hypothesis. For one, Snow would need a way to trick the data such that the allocation of clean and dirty water to people was not associated with the other determinants of cholera mortality, such as hygiene and poverty. He just would need for someone or something to be making this treatment assignment for him.

Fortunately for Snow, and the rest of London, that someone or something existed. In the London of the 1800s, there were many different water companies serving different areas of the city. Some were served by more than one company. Several took their water from the Thames, which was heavily polluted by sewage. The service areas of such companies had much higher rates of cholera. The Chelsea water company was an exception, but it had an exceptionally good filtration system. That's when Snow had a major insight. In 1849, Lambeth water company moved the intake point upstream along the Thames, above the main sewage discharge point, giving its customers purer water. Southwark and Vauxhall water company, on the other hand, left their intake point downstream from where the sewage discharged. Insofar as the kinds of people that each company serviced were approximately the same, then comparing the cholera rates between the two houses could be the experiment that Snow so desperately needed to test his hypothesis.

Snow's Table IX			
Company name	Number of houses	Cholera deaths	Deaths per 10,000 houses
Southwark and Vauxhall	40,046	1,263	315
Lambeth	26,107	98	37

Snow wrote up his results in a document with many tables and a map showing the distribution of cholera cases around the city - one of the first statistical maps, and one of the most famous. Table 9, above, shows the main results. Southwark and Vauxhall, what I call the treatment case, had 1,263 cholera deaths, which is 315 per 10,000 houses. Lambeth, the control, had only 98, which is 37 per 10,000 houses. Snow spent the majority of his time in the write

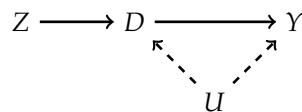
up tediously documenting the similarities between the groups of domiciles serviced by the two companies in order to rule out the possibility that some other variable could be both correlated with Southwark and Vauxhall and associated with miasmis explanations. He was convinced – cholera was spread through the water supply, not the air. Of this table, Freedman [1991] the statistician wrote:

“As a piece of statistical technology, [Snow’s Table IX] is by no means remarkable. But the story it tells is very persuasive. The force of the argument results from the clarity of the prior reasoning, the bringing together of many different lines of evidence, and the amount of shoe leather Snow was willing to use to get the data. Snow did some brilliant detective work on nonexperimental data. What is impressive is not the statistical technique but the handling of the scientific issues. He made steady progress from shrewd observation through case studies to analyze ecological data. In the end, he found and analyzed a natural experiment.”

The idea that the best instruments come from shoelather is echoed in Angrist and Krueger [2001] when the authors note that the best instruments come from in-depth knowledge of the institutional details of some program or intervention.

Instrumental variables DAG

To understand the instrumental variables estimator, it is helpful to start with a DAG. This DAG shows a chain of causal effects that contains all the information needed to understand the instrumental variables strategy. First, notice the backdoor path between D and Y : $D \leftarrow u \rightarrow Y$. Furthermore, note that u is unobserved by the econometrician which causes the backdoor path to remain open. If we have this kind of *selection on unobservables*, then there does not exist a conditioning strategy that will satisfy the backdoor criterion (in our data). But, before we throw up our arms, let’s look at how Z operates through these pathways.



First, there is a mediated pathway from Z to Y via D . When Z varies, D varies, which causes Y to change. But, even though Y is varying when Z varies, notice that Y is only varying *because* D has varied. You sometimes hear people describe this as the “only through” assumption. That is, Z affects Y “only through” D .

Imagine this for a moment though. Imagine D consists of people making choices. Sometimes these choices affect Y , and sometimes

these choices merely reflect changes in Y via changes in U . But along comes some shock, Z , which induces *some* but not all of the people in D to make different decisions. What will happen?

Well, for one, when those people's decisions change, Y will change too, because of the causal effect. But, notice, all of the correlation between D and Y in that situation will reflect the causal effect. The reason being, D is a collider along the backdoor path between Z and Y .

But I'm not done with this metaphor. Let's assume that in this D variable, with all these people, only some of the people change their behavior because of D . What then? Well, in that situation, Z is causing a change in Y for just a subset of the population. If the instrument only changes the behavior of women, for instance, then the causal effect of D on Y will only reflect the causal effect of *female choices*, not males.

There's two ideas inherent in the previous paragraph that I want to emphasize. First, if there are heterogeneous treatment effects (e.g., males affect Y differently than females), then our Z shock only identified some of the causal effect of D on Y . And that piece of the causal effect may only be valid for the female population whose behavior changed in response to Z ; it may not be reflective of how male behavior would affect Y . And secondly, if Z is only inducing some of the change in Y via only a fraction of the change in D , then it's almost as though we have less data to identify that causal effect than we really have.

Here we see two of the difficulties in both interpreting instrumental variables, as well as identifying a parameter with it. Instrumental variables only identifies a causal effect for any group of units whose behaviors are changed as a result of the instrument. We call this the causal effect of the *complier* population; in our example, only females "complied" with the instrument, so we only know its effect for them. And secondly, instrumental variables are typically going to have larger standard errors, and as such, will fail to reject in many instances if for no other reason than because they are under-powered.

Moving along, let's return to the DAG. Notice that we drew the DAG such that Z has no connection to U . Z is independent of U . That is called the "exclusion restriction" which we will discuss in more detail later. But briefly, the IV estimator assumes that Z is independent of the variables that determine Y *except* for D .

Secondly, Z is correlated with D and because of its correlation with D (and D 's effect on Y), Z is correlated with Y but only through its effect on D . This relationship between Z and D is called the "first stage", named that because of the two stage least squares estimator, which is a kind of IV estimator. The reason it is only correlated with

Y via D is because D is a collider along the path $Z \rightarrow D \leftarrow u \rightarrow Y$.

How do you know when you have a good instrument? One, it will require a DAG - either an explicit one, or an informal one. You can only identify a causal effect using IV if you can theoretically and logically defend the exclusion restriction, since the exclusion restriction is an untestable assumption technically. That defense requires theory, and since some people aren't comfortable with theoretical arguments like that, they tend to eschew the use of IV. More and more, applied microeconomists are skeptical of IV for this reason.

But, let's say you think you do have a good instrument. How might you defend it as such to someone else? A necessary but not a sufficient condition for having an instrument that can satisfy the exclusion restriction is if people are confused when you tell them about the instrument's relationship to the outcome. Let me explain. No one is going to be confused when you tell them that you think family size will reduce female labor supply. They don't need a Becker model to convince them that women who have more children probably work less than those with fewer children. It's common sense. But, what would they think if you told them that mothers whose first two children were the same gender worked less than those whose children had a balanced sex ratio? They would probably give you a confused look. What does the gender composition of your children have to do with whether a woman works?

It doesn't – it only matters, in fact, if people whose first two children are the same gender decide to have a third child. Which brings us back to the original point – people buy that family size can cause women to work less, but they're confused when you say that women work less when their first two kids are the same gender. But if when you point out to them that the two children's gender induces people to have larger families than they would have otherwise, the person "gets it", then you might have an excellent instrument.

Instruments are, in other words, jarring. They're jarring precisely because of the exclusion restriction – these two things (gender composition and work) don't seem to go together. If they did go together, it would likely mean that the exclusion restriction was violated. But if they don't, then the person is confused, and that is at minimum a possible candidate for a good instrument. This is the common sense explanation of the "only through" assumption.

The following two sections differ from one another in the following sense: the next section makes the traditional assumption that all treatment effects are constant for all units. When this is assumed, then the parameter estimated through an IV methodology equals the ATE equals the ATT equals the ATU. The variance will still be

larger, because IV still only uses part of the variation in D , but the compliers are identical to the non-compliers so the causal effect for the compliers is the same as the causal effect for all units.

The section after the next one is explicitly based on the potential outcomes model. It assumes the more general case where each unit has a unique treatment effect. If each unit can have a different effect on Y , then the causal effect itself is a random variable. We've called this heterogeneous treatment effects. It is in this situation that the complier qualification we mentioned earlier matters, because if we are only identifying a causal effect for just a subset of the column of causal effects, then we are only estimating the treatment effects associated with the compliers themselves. This estimand is called the *local average treatment effect* (LATE), and it adds another wrinkle to your analysis. If the compliers' own average treatment effects are radically different from the rest of the population, then the LATE estimand may not be very informative. Heck, under heterogeneous treatment effects, there's nothing stopping the *sign* of the LATE to be different than the sign of the ATE!

For this reason, we want to think long and hard about what our IV estimate means under heterogeneous treatment effects, because policy-makers will be implementing a policy, not based on assigning Z but rather based on assigning D . And as such, both compliers and non-compliers will matter for the policy-makers, yet IV only identifies the effect for one of these. Hopefully this will become clearer as we progress.

Homogenous treatment effects and 2SLS

Instrumental variables methods are typically used to address omitted variable bias, measurement error, and simultaneity. For instance, quantity and price is determined by the intersection of supply and demand, so any observational correlation between price and quantity is uninformative about the unique elasticities associated with supply or demand curves. Wright understood this, which was why he investigated the problem so intensely.

We begin by assuming homogenous treatment effects. Homogeneous treatment effects assumes that the treatment effect is the same for every unit. This is the traditional econometric pedagogy and not based explicitly on the potential outcomes notation.

Let's start by illustrating the omitted variable bias problem again. Assume the classical labor problem where we're interested in the causal effect of schooling on earnings, but schooling is endogenous

because of unobserved ability. Let the true model of earnings be:

$$Y_i = \alpha + \delta S_i + \gamma A_i + \varepsilon_i$$

where Y is the log of earnings, S is schooling measured in years, A is individual “ability”, and ε is an error term uncorrelated with schooling or ability. The reason A is unobserved is simply because the surveyor either forgot to collect it or couldn’t collect it and therefore it’s missing from your dataset.¹⁰⁹ For instance, the CPS tells us nothing about respondents’ family background, intelligent, motivation or non-cognitive ability. Therefore, since ability is unobserved, we have the following equation instead:

$$Y_i = \alpha + \delta S_i + \eta_i$$

where η_i is a composite error term equalling $\gamma A_i + \varepsilon_i$. We assume that schooling is correlated with ability, so therefore it is correlated with η_i , making it endogenous in the second, shorter regression. Only ε_i is uncorrelated with the regressors, and that is by definition.

We know from the derivation of the least squares operator that the estimated value of $\hat{\delta}$ is:

$$\hat{\delta} = \frac{C(Y, S)}{V(S)} = \frac{E[YS] - E[Y]E[S]}{V(S)}$$

Plugging in the true value of Y (from the longer model), we get the following:

$$\begin{aligned}\hat{\delta} &= \frac{E[\alpha S + S^2\delta + \gamma SA + \varepsilon S] - E(S)E[\alpha + \delta S + \gamma A + \varepsilon]}{V(S)} \\ &= \frac{\delta E(S^2) - \delta E(S)^2 + \gamma E(AS) - \gamma E(S)E(A) + E(\varepsilon S) - E(S)E(\varepsilon)}{V(S)} \\ &= \delta + \gamma \frac{C(AS)}{V(S)}\end{aligned}$$

If $\gamma > 0$ and $C(AS) > 0$, then $\hat{\delta}$, the coefficient on schooling, is upward biased. And that is probably the case given that it’s likely that ability and schooling are positively correlated.

Now, consistent with the IV DAG we discussed earlier, suppose there exists a variable, Z_i , that is correlated with schooling. We can use this variable, as I’ll now show, to estimate δ . First, calculate the covariance of Y and Z :

$$\begin{aligned}C(Y, Z) &= C(\alpha\delta S + \gamma A + \varepsilon, Z) \\ &= E[(\alpha + \delta S + \gamma A + \varepsilon), Z] - E(S)E(Z) \\ &= \{\alpha E(Z) - \alpha E(Z)\} + \delta\{E(SZ) - E(S)E(Z)\} + \gamma\{E(AZ) - E(A)E(Z)\} + \{E(\varepsilon Z) - E(\varepsilon)E(Z)\} \\ &= \delta C(S, Z) + \gamma C(A, Z) + C(\varepsilon, Z)\end{aligned}$$

¹⁰⁹ Unobserved ability doesn’t mean it’s literally unobserved, in other words. It could be just missing from your dataset, and therefore is unobserved *to you*.

Notice that the parameter of interest, δ is on the right hand side. So how do we isolate it? We can estimate it with the following:

$$\hat{\delta} = \frac{C(Y, Z)}{C(S, Z)}$$

so long as $C(A, Z) = 0$ and $C(\varepsilon, Z) = 0$.

These zero covariances are the statistical truth contained in the IV DAG from earlier. If ability is independent of Z , then this second covariance is zero. And if Z is independent of the structural error term, ε , then it too is zero. This, you see, is what is meant by the “exclusion restriction”: the instrument must be independent of both parts of the composite error term.

But the exclusion restriction is only a necessary condition for IV to work; it is not a sufficient condition. After all, if all we needed was exclusion, then we could use a random number generator for an instrument. Exclusion is not enough. We also need the instrument to be *highly correlated* with the endogenous variable. And the higher the better. We see that here because we are dividing by $C(S, Z)$, so it necessarily requires that this covariance be non-zero.

The numerator in this simple ratio is sometimes called the “reduced form”, while the denominator is called the “first stage”. These terms are somewhat confusing, particularly the former as “reduced form” means different things to different people. But in the IV terminology, it is that relationship between the instrument and the outcome itself. The first stage is less confusing, as it gets its name from the two stage least squares estimator, which we’ll discuss next.

When you take the probability limit of this expression, then assuming $C(A, Z) = 0$ and $C(\varepsilon, Z) = 0$ due to the exclusion restriction, you get

$$\text{plim } \hat{\delta} = \delta$$

But if Z is not independent of η (either because it’s correlated with A or ε), and if the correlation between S and Z is weak, then $\hat{\delta}$ becomes severely biased.

Two stage least squares One of the more intuitive instrumental variables estimators is the two-stage least squares (2SLS). Let’s review an example to illustrate why I consider it helpful for explaining some of the IV intuition. Suppose you have a sample of data on Y , S and Z . For each observation i , we assume the data are generated according to:

$$\begin{aligned} Y_i &= \alpha + \delta S_i + \varepsilon_i \\ S_i &= \gamma + \beta Z_i + \epsilon_i \end{aligned}$$

where $C(Z, \varepsilon) = 0$ and $\beta \neq 0$. Now using our IV expression, and using the result that $\sum_{i=1}^n (x_i - \bar{x}) = 0$, we can write out the IV estimator as:

$$\begin{aligned}\hat{\delta} &= \frac{C(Y, Z)}{C(S, Z)} \\ &= \frac{\frac{1}{n} \sum_{i=1}^n (Z_i - \bar{Z})(Y_i - \bar{Y})}{\frac{1}{n} \sum_{i=1}^n (Z_i - \bar{Z})(S_i - \bar{S})} \\ &= \frac{\frac{1}{n} \sum_{i=1}^n (Z_i - \bar{Z})Y_i}{\frac{1}{n} \sum_{i=1}^n (Z_i - \bar{Z})S_i}\end{aligned}$$

When we substitute the true model for Y , we get the following:

$$\begin{aligned}\hat{\delta} &= \frac{\frac{1}{n} \sum_{i=1}^n (Z_i - \bar{Z})\{\alpha + \delta S + \varepsilon\}}{\frac{1}{n} \sum_{i=1}^n (Z_i - \bar{Z})S_i} \\ &= \delta + \frac{\frac{1}{n} \sum_{i=1}^n (Z_i - \bar{Z})\varepsilon_i}{\frac{1}{n} \sum_{i=1}^n (Z_i - \bar{Z})S_i} \\ &= \delta + \text{"small if } n \text{ is large"}$$

So, let's return to our first description of $\hat{\delta}$ as the ratio of two covariances. With some simple algebraic manipulation, we get the following:

$$\begin{aligned}\hat{\delta} &= \frac{C(Y, Z)}{C(S, Z)} \\ &= \frac{\frac{C(Z, Y)}{V(Z)}}{\frac{C(Z, S)}{V(Z)}}\end{aligned}$$

where the denominator is equal to $\hat{\beta}$.¹¹⁰ We can rewrite $\hat{\beta}$ as:

¹¹⁰ That is, $S_i = \gamma + \beta Z_i + \varepsilon_i$

$$\begin{aligned}\hat{\beta} &= \frac{C(Z, S)}{V(Z)} \\ \hat{\beta}V(Z) &= C(Z, S)\end{aligned}$$

Then we rewrite the IV estimator and make a substitution:

$$\begin{aligned}\hat{\delta}_{IV} &= \frac{C(Z, Y)}{C(Z, S)} \\ &= \frac{\hat{\beta}C(Z, Y)}{\hat{\beta}C(Z, S)} \\ &= \frac{\hat{\beta}C(Z, Y)}{\hat{\beta}^2 V(Z)} \\ &= \frac{C(\hat{\beta}Z, Y)}{V(\hat{\beta}Z)}\end{aligned}$$

Recall that $S = \gamma + \beta Z + \epsilon$; $\hat{\delta} = \frac{C(\hat{\beta}ZY)}{V(\hat{\beta}Z)}$ and let $\hat{S} = \hat{\gamma} + \hat{\beta}Z$. Then the 2SLS estimator is:

$$\begin{aligned}\hat{\delta}_{IV} &= \frac{C(\hat{\beta}Z, Y)}{V(\hat{\beta}Z)} \\ &= \frac{C(\hat{S}, Y)}{V(\hat{S})}\end{aligned}$$

I will now show that $\hat{\beta}C(Y, Z) = C(\hat{S}, Y)$, and leave it to you to show that $V(\hat{\beta}Z) = V(\hat{S})$.

$$\begin{aligned}C(\hat{S}, Y) &= E[\hat{S}Y] - E[\hat{S}]E[Y] \\ &= E(Y[\hat{\gamma} + \hat{\beta}Z]) - E(Y)E(\hat{\gamma} + \hat{\beta}Z) \\ &= \hat{\gamma}E(Y) + \hat{\beta}E(YZ) - \hat{\gamma}E(Y) - \hat{\beta}E(Y)E(Z) \\ &= \hat{\beta}[E(YZ) - E(Y)E(Z)] \\ C(\hat{S}, Y) &= \hat{\beta}C(Y, Z)\end{aligned}$$

Now let's return to something I said earlier – learning 2SLS can help you better understand the intuition of instrumental variables more generally. What does this mean exactly? It means several things. First, the 2SLS estimator used only the fitted values of the endogenous regressors for estimation. These fitted values were based on all variables used in the model, *including the excludable instrument*. And as all of these instruments are exogenous in the structural model, what this means is that the fitted values themselves have become exogenous too. Put differently, we are using only the variation in schooling that is *exogenous*. So that's kind of interesting, as now we're back in a world where we are identifying causal effects.

B

But, now the less exciting news. This exogenous variation in S driven by the instruments is only a subset of the total variation in the variable itself. Or put differently, IV reduces the variation in the data, so there is less information available for identification, and what little variation we have left comes from the *complier* population only. Hence the reason in large samples we are estimating the LATE – that is, the causal effect for the complier population, where a complier is someone whose behavior was altered by the instrument.

Example 1: Meth and Foster Care

As before, I feel that an example will help make this strategy more concrete. To illustrate, I'm going to review one of my papers with Keith Finlay examining the effect of methamphetamine abuse on child abuse and foster care admissions [Cunningham and Finlay,

2012]. It has been claimed that substance abuse, notably drug use, has a negative impact on parenting, such as neglect, but as these all occur in equilibrium, it's possible that the correlation is simply reflective of selection bias. In other words, perhaps households with parents who abuse drugs would've had the same negative outcomes had the parents not used drugs. After all, it's not like people are flipping coins when deciding to use meth. So let me briefly give you some background to the study so that you better understand the data generating process.

First, d-methamphetamine is like poison to the mind and body when abused. Effects meth abuse increase energy and alertness, decreased appetite, intense euphoria, impaired judgment, and psychosis. Second, the meth epidemic, as it came to be called, was geographically concentrated initially on the west coast before gradually making its way eastward over the 1990s.

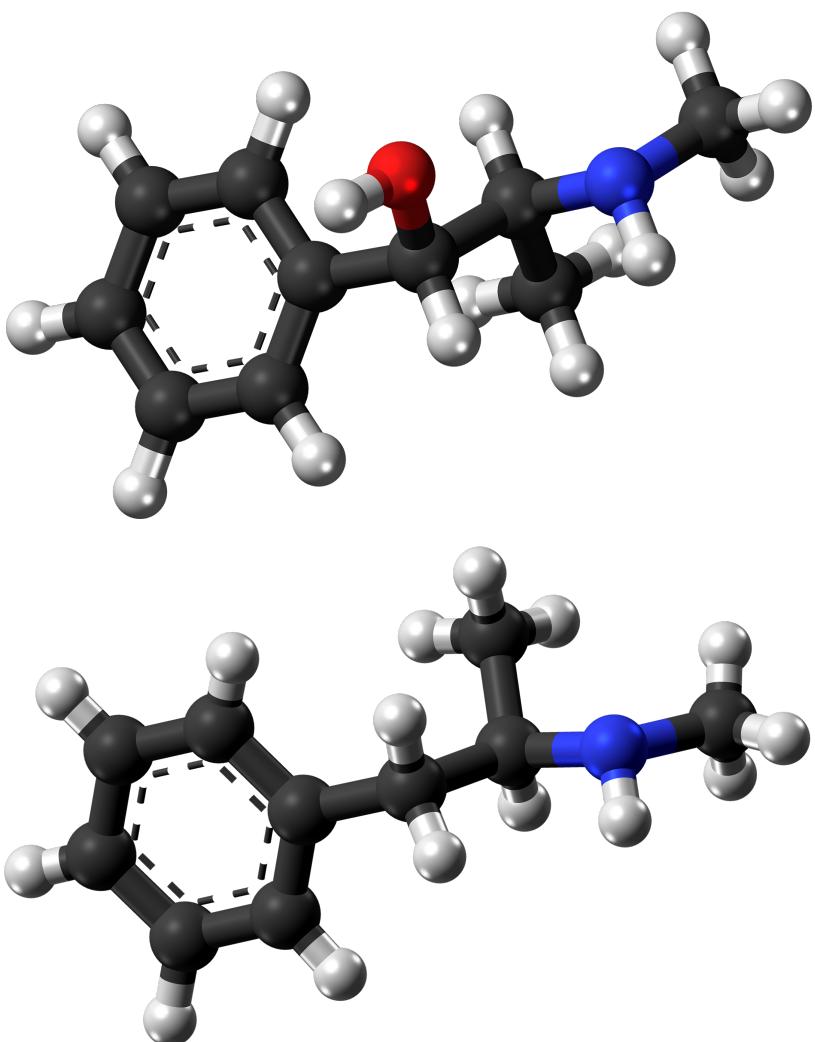
What made this study possible, though, was meth's production process. Meth is synthesized from a reduction of ephedrine or pseudoephedrine, which is also the active ingredient in many cold medications, such as the behind-the-counter Sudafed. It is also worth noting that that key input (precursor) experienced a bottleneck in production. In 2004, nine factories manufactured the bulk of the world supply of ephedrine and pseudoephedrine. The DEA correctly noted that if they could regulate access to ephedrine and pseudoephedrine, then they could effectively interrupt the production of d-methamphetamine, and in turn, reduce meth abuse and its associated social harms.

To understand this, it may be useful to see the two chemical molecules side by side. While the actual process of production is more complicated than this, the chemical reduction is nonetheless straightforward: start with ephedrine or pseudoephedrine, remove the hydroxyl group, add back the hydrogen. This gives you d-methamphetamine (see Figure 64).

So, with input from the DEA, Congress passed the Domestic Chemical Diversion Control Act in August 1995 which provided safeguards by regulating the distribution of products that contained ephedrine as the only medicinal ingredient. But the new legislation's regulations applied to ephedrine, not pseudoephedrine, and since the two precursors were nearly identical, traffickers quickly substituted from ephedrine to pseudoephedrine. By 1996, pseudoephedrine was found to be the primary precursor in almost half of meth lab seizures.

Therefore, the DEA went back to Congress, seeking greater control over pseudoephedrine products. And the Comprehensive Methamphetamine Control Act of 1996 went into effect between October

Figure 64: Pseudoephedrine (top) vs d-methamphetamine (bottom)

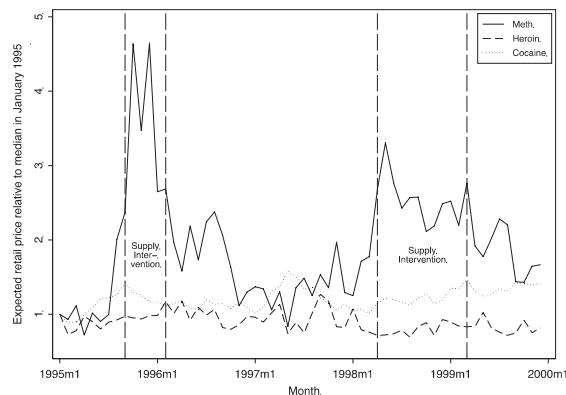


and December 1997. This Act required distributors of all forms of pseudoephedrine to be subject to chemical registration. Dobkin and Nicosia [2009] argued that these precursor shocks may very well have been the largest supply shocks in the history of drug enforcement.

The effect of the two interventions were dramatic. The first supply intervention caused retail (street) prices (adjusted for purity, weight and inflation) to more than quadruple. The second more like 2-3 times its longrun trend. See Figure 65.

We are interested in the causal effect of meth abuse on child abuse, and so our first stage is necessarily a proxy for meth abuse – the number of people entering treatment who listed meth as one of the substances they used in their last substance abuse episode. As I said

FIGURE 3
Ratio of Median Monthly Expected Retail Prices of Meth, Heroin, and Cocaine Relative to Their Respective Values in January 1995, STRIDE, 1995–1999



before, since pictures speak a thousand words, I'm going to show you pictures of both the first stage and the reduced form. Why do I do this instead of going directly to the tables of coefficients? Because quite frankly, you are more likely to find those estimates believable if you can see evidence for the first stage and the reduced form in the raw data itself.¹¹¹

In Figure 66, we show the first stage and you can see several things. All of these data come from the Treatment Episode Data Set (TEDS), which is all people going into treatment for substance abuse for federally funded clinics. Patients list the last three substances used in the most recent "episode". We mark anyone who listed meth, cocaine or heroin as counts by month and state. Here we aggregate to the national level in Figure 66. You can see evidence for the effect the two interventions had on meth flows, particularly the ephedrine intervention. Self-admitted meth admissions dropped significantly, as did total meth admissions, but there's no effect on cocaine or heroin. The effect of the pseudoephedrine is not as dramatic, but it appears to cause a break in trend as the growth in meth admissions slows during this period of time.

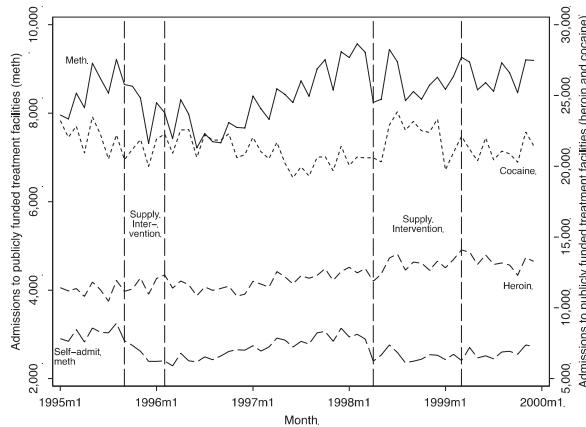
In Figure 67, we graphically show the reduced form. That is, the effect of the price shocks on foster care admissions. Consistent with what we found in our first stage graphic, the ephedrine intervention in particular had a profoundly negative effect on foster care admissions. They fell from around 8,000 children removed per month to around 6,000, then began rising again. The second intervention also had an effect, though it appears to be milder. The reason we believe that the second intervention had a more modest effect than the first is because (1) the effect on price as we saw earlier was about half the

Figure 65: Figure 3 from Cunningham and Finlay [2012] showing changing street prices following both supply shocks.

¹¹¹ While presenting figures of the first stage and reduced form isn't mandatory in the way that it is for regression discontinuity, it is nonetheless very commonly done. Ultimately, it is done because seeing is believing.

FIGURE 5

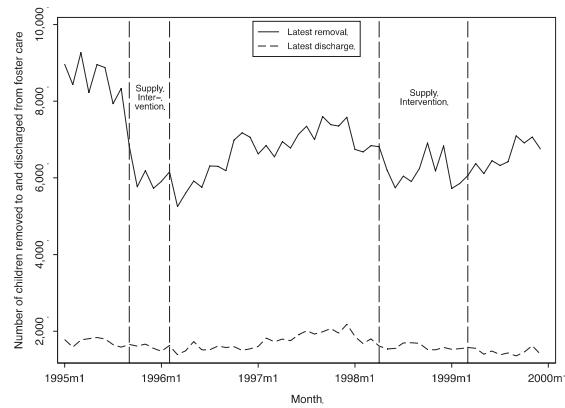
Total Admissions to Publicly Funded Treatment Facilities by Drug and Month, Selected States,
Whites, TEDS, Seasonally Adjusted, 1995–1999



size of the first intervention, and (2) domestic meth production was being replaced by Mexican imports of d-meth over the late 1990s. Thus, by the end of the 1990s, domestic meth production played a smaller role in total output, hence why the effect on price and admissions was probably smaller.

FIGURE 4

Number of Children Removed to and Discharged from Foster Care in a Set of Five States by Month, AFCARS, Seasonally Adjusted, 1995–1999



In Figure 68, we reproduce Table 3 from my article with Keith. There are a few pieces of key information that all IV tables should have. First, there is the OLS regression. As the OLS regression suffers from endogeneity, we want the reader to see it so that they what

Figure 66: Figure 5 from Cunningham and Finlay [2012] showing first stage.

Figure 67: Figure 4 from Cunningham and Finlay [2012] showing reduced form effect of interventions on children removed from families and placed into foster care.

to compare the IV model with. Let's focus on column 1 where the dependent variable is total entry into foster care. We find no effect, interestingly, of meth onto foster care.

Covariates	Log Latest Entry into Foster Care		Log Latest Entry via Parental Incarceration		Log Latest Entry via Child Neglect	
	OLS (1)	2SLS (2)	OLS (3)	2SLS (4)	OLS (5)	2SLS (6)
Log self-referred meth treatment rate	0.01 (0.02)	1.54*** (0.59)	0.23*** (0.05)	-0.38 (0.32)	0.03 (0.02)	1.03** (0.41)
Unemployment rate	-0.06** (0.02)	-0.00 (0.05)	-0.04 (0.06)	-0.04 (0.06)	-0.07*** (0.02)	-0.03 (0.04)
Cigarette tax per pack	-0.01 (0.10)	0.02 (0.17)	-2.02*** (0.42)	-1.96*** (0.42)	0.15 (0.12)	0.16 (0.16)
Log alcohol treatment rate	-0.04 (0.03)	-1.26*** (0.46)	-0.37 (0.09)	0.13 (0.28)	-0.05 (0.03)	-0.85*** (0.32)
Log population 0–19 year old	3.68 (2.59)	2.25 (3.60)	-42.61* (22.74)	-40.43* (22.24)	2.12 (2.66)	1.28 (3.21)
Log population 15–49 year old	-15.48*** (5.44)	-10.61* (6.19)	-27.20 (22.20)	-32.24 (21.35)	-8.93* (5.11)	-5.66 (5.52)
Month-of-year fixed effects	x	x	x	x	x	x
State fixed effects	x	x	x	x	x	x
State linear time trends	x	x	x	x	x	x
<i>First stage</i>						
Price deviation instrument		-0.0005*** (0.0001)		-0.0009*** (0.0002)		-0.0005*** (0.0001)
F-statistic for IV in first stage		17.60		25.99		18.78
R ²	0.864		0.818		0.855	
N	1,343		1,068		1,317	
	Log Latest Entry via Parental Drug Use		Log Latest Entry via Physical Abuse		Log Number of Exits from Foster Care	
	OLS (7)	2SLS (8)	OLS (9)	2SLS (10)	OLS (11)	2SLS (12)
Log self-referred meth treatment rate	0.21*** (0.04)	-0.20 (0.34)	0.04 (0.03)	1.49** (0.62)	0.06* (0.03)	-0.14 (0.28)
Unemployment	-0.17*** (0.05)	-0.18*** (0.05)	-0.11*** (0.04)	-0.05 (0.06)	-0.02 (0.03)	-0.03 (0.03)
Cigarette tax per pack	-2.80*** (0.37)	-2.80*** (0.36)	0.17 (0.14)	0.20 (0.19)	-1.05*** (0.15)	-1.05*** (0.15)
Log alcohol treatment rate	-0.24*** (0.07)	0.10 (0.28)	-0.01 (0.05)	-1.16** (0.49)	-0.04 (0.04)	0.12 (0.22)
Log population 0–19 year old	-13.30 (17.74)	-10.59 (18.22)	0.81 (3.73)	-0.44 (4.18)	9.50*** (3.60)	9.69*** (3.51)
Log population 15–49 year old	-0.71 (33.63)	-6.01 (34.71)	-8.74 (6.83)	-4.01 (7.01)	-20.22*** (5.39)	-20.90*** (5.33)
Month-of-year fixed effects	x	x	x	x	x	x
State fixed effects	x	x	x	x	x	x
State linear time trends	x	x	x	x	x	x
<i>First stage</i>						
Price deviation instrument		-0.0007*** (0.0001)		-0.0005*** (0.0001)		-0.0005*** (0.0001)
F-statistic for IV in first stage		24.45		18.29		17.70
R ²	0.90		0.80		0.84	
N	1,161		1,293		1,318	

Figure 68: Table 3 Cunningham and Finlay [2012] showing OLS and 2SLS estimates of meth on foster care admissions.

The second piece of information that one should report in a 2SLS table is the first stage itself. We report the first stage at the bottom of each even numbered column. As you can see, for each one unit deviation in price from its longrun trend, meth admissions into treatment (our proxy) fell by -0.0005 log points. This is highly significant at the 1% level, but we check for the strength of the instrument using the F statistic [Staiger and Stock, 1997].¹¹² We have an F statistic of 17.6, which suggests that our instrument is strong enough for identification.

Finally, the 2SLS estimate of the treatment effect itself. Notice using only the exogenous variation in log meth admissions, and

¹¹² In a sense, I am probably getting ahead of myself as we technically haven't introduced weak instrument tests. But I wanted to walk you through an IV paper before getting too far into the weeds. We will circle back around and discuss weak instruments later, but for now know that Staiger and Stock [1997] suggested that weak instruments were a problem when an F test on the excludability of the instrument from the first stage was less than 10. That paper was not the last word. See Stock and Yogo [2005] if you're interested in precise, quantitative definitions of weak instruments.

assuming the exclusion restriction holds in our model, we are able to isolate a causal effect of log meth admissions on log aggregate foster care admissions. As this is a log-log regression, we can interpret the coefficient as an elasticity. We find that a 10% increase in meth admissions for treatment appears to cause around a 15% increase in children removed from their homes and placed into foster care. This effect is both large and precise. And notice, it was not detectable otherwise (the coefficient was zero).

Why are they being removed? Our data (AFCARS) lists several channels: parental incarceration, child neglect, parental drug use, and physical abuse. Interestingly, we do not find any effect of parental drug use or parental incarceration, which is perhaps somewhat counterintuitive. Their signs are negative and their standard errors are large. Rather, we find effects of meth admissions on removals for physical abuse and neglect. Both are elastic (i.e., > 1).

What did we learn from this paper? Well, we learned two kinds of things. First, we learned how a contemporary piece of applied microeconomics goes about using instrumental variables to identify causal effects. We saw the kinds of graphical evidence mustered, the way in which knowledge about the natural experiment and the policies involved helped the authors argue for the exclusion restriction (since it cannot be tested), and the kind of evidence presented from 2SLS, including the first stage tests for weak instruments. Hopefully seeing a paper at this point was helpful. But the second thing we learned concerned the actual study itself. We learned that for the group of meth users whose behavior was changed as a result of rising real prices of a pure gram of d-methamphetamine (i.e., the complier subpopulation), their meth use was causing child abuse and neglect that was so severe that it merited removing their children and placing those children into foster care. If you were only familiar with [Dobkin and Nicosia \[2009\]](#), who found no effect of meth on crime using county level data from California and only the 1997 ephedrine shock, you might incorrectly conclude that there are no social costs associated with meth abuse. But, while meth does not appear to cause crime, it does appear to harm the children of meth users and place strains on the foster care system.

Example 2: Compulsory Schooling and Weak Instruments

I am not trying to smother you with papers. But before we move back into the technical material itself, I'd like to discuss one more paper. This paper is interesting and important in and of itself, but even putting that aside, it will also help you better understand the weak instrument literature which followed.

As we've said since the beginning, with example of example, there is a very long tradition in labor economics of building models that can credibly identify the returns to schooling. This goes back to Becker [1994] and the Labor workshop at Columbia that Becker ran for years with Jacob Mincer. This has been an important task given education's growing importance in the distribution of income and wealth in the latter 20th century due to the increasing returns to skill in the marketplace [Juhn et al., 1993].

One of the more seminal papers in instrumental variables for the modern period is Angrist and Krueger [1991]. The idea is simple and clever; a quirk in the United States educational system is that a child is chosen for one grade based on when their birthday is. For a long time, that cutoff was late December. If a child was born on or before December 31st, then they were assigned to the first grade. But if their birthday was on or after January 1st, they were assigned to kindergarten. Thus these two people – one born on December 31st and one born on January 1st – were exogenously assigned different grades.

Now there's nothing necessarily relevant here because if they always stay in school for the duration of time necessary to get a high school degree, then that arbitrary assignment of start date won't affect high school completion. It'll only affect *when* they get that high school degree. But this is where the quirk gets interesting. For most of the 20th century, the US had compulsory schooling laws which forced a person to remain in high school until they reached age 16. After they hit age 16, they could legally drop out. Figure 69 explains visually their instrumental variable.

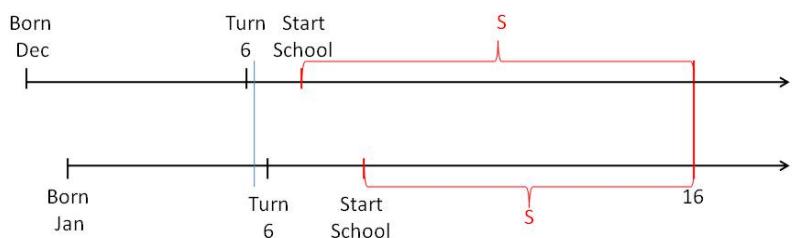


Figure 69: Angrist and Krueger [1991] explanation of their instrumental variable.

Angrist and Krueger had the insight that that small quirk was exogenously assigning more schooling to people born later in the year. The person born in December would reach age 16 with more education than the person born in January, in other words. Thus, the authors had exogenous variation in schooling.¹¹³

In Figure 70, Angrist and Krueger [1991] visually show the reader the first stage, and it is really interesting. There's a clear pattern -

¹¹³ Notice how similar their idea was to regression discontinuity. That's because IV and RDD are conceptually very similar strategies.

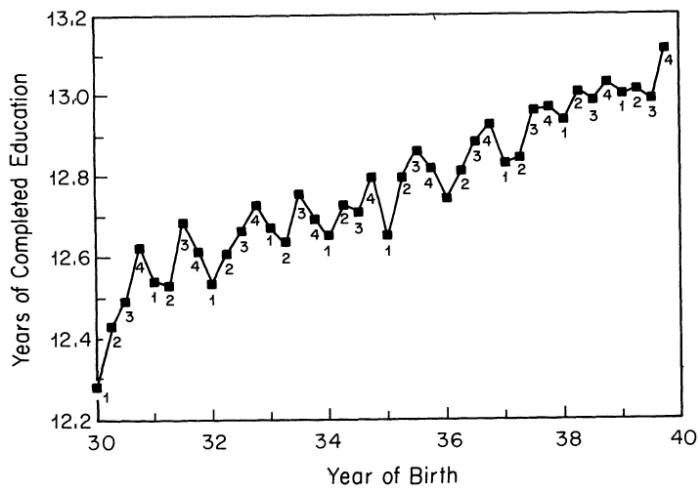


Figure 70: Angrist and Krueger [1991] first stage relationship between quarter of birth and schooling.

3rd and 4th quarter birth days have more schooling than 1st and 2nd quarter births on average. That relationship gets weaker as we move into later cohorts, but that is probably because for later cohorts, the price on higher levels of schooling was rising so much that fewer and fewer people were dropping out before finishing their high school degree.

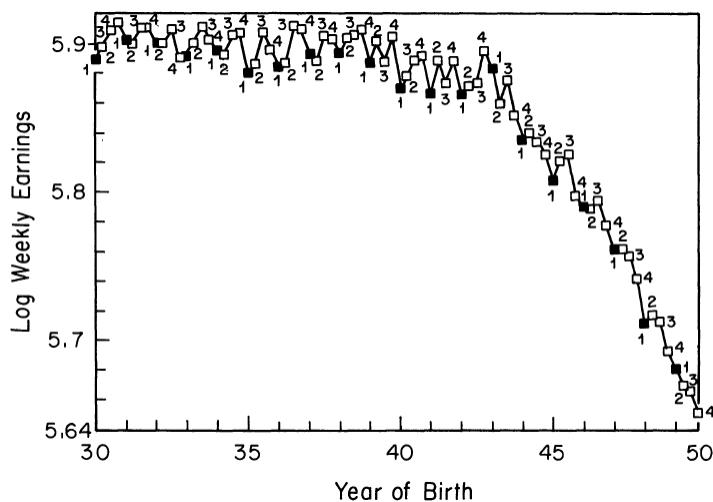


Figure 71: Angrist and Krueger [1991] reduced form visualization of the relationship between quarter of birth and log weekly earnings.

Figure 71 shows the reduced form visually. That is, here we see a simple graph showing the relationship between quarter of birth and log weekly earnings.¹¹⁴ You have to squint your eye a little bit, but you can see the pattern – all along the top of the jagged path are 3s

¹¹⁴ I know, I know. No one has ever accused me of being subtle. But it's an important point - a picture speaks a thousand words. If you can communicate your first stage and reduced form in pictures, you always should, as it will really captivate the reader's attention and be far more compelling than a simple table of coefficients ever could.

and 4s, and all along the bottom of the jagged path are 1s and 2s. Not always, but it's correlated.

Let's take a sidebar. Remember what I said about how instruments have a certain ridiculousness to them? That is, you know you have a good instrument if the instrument itself doesn't seem relevant for explaining the outcome of interest because *that's what the exclusion restriction implies*. Why would quarter of birth affect earnings? It doesn't make any obvious, logical sense why it should. But, if I told you that people born later in the year got more schooling than those with less *because of compulsory schooling*, then the relationship between the instrument and the outcome snaps into place. The only reason we can think of as to why the instrument would affect earnings is if the instrument was operating through schooling. Instruments only explain the outcome, in other words, when you understand their effect on the endogenous variable.¹¹⁵

Angrist and Krueger use three dummies as their instruments: a dummy for first quarter, a dummy for second quarter and a dummy for third quarter. Thus the omitted category is the fourth quarter, which is the group that gets the most schooling. Now ask yourself this: if we regressed years of schooling onto those three dummies, what should the signs and magnitudes be? That is, what would we expect the relationship between the first quarter (compared to the fourth quarter) and schooling? Let's look at their first stage results and see if it matched your intuition (Figure 72).

¹¹⁵ This is why I chose those particular Chance the Rapper lyrics as this chapter's epigraph. There's no reason why making "Sunday Candy" would keep Chance from going to hell. Without knowing the first stage, it makes no obvious sense!

Figure 72: Angrist and Krueger [1991] first stage for different outcomes.

Outcome variable	Birth cohort	Mean	Quarter-of-birth effect ^a			F-test ^b [P-value]
			I	II	III	
Total years of education	1930–1939	12.79	−0.124 (0.017)	−0.086 (0.017)	−0.015 (0.016)	24.9 [0.0001]
	1940–1949	13.56	−0.085 (0.012)	−0.035 (0.012)	−0.017 (0.011)	18.6 [0.0001]
High school graduate	1930–1939	0.77	−0.019 (0.002)	−0.020 (0.002)	−0.004 (0.002)	46.4 [0.0001]
	1940–1949	0.86	−0.015 (0.001)	−0.012 (0.001)	−0.002 (0.001)	54.4 [0.0001]
Years of educ. for high school graduates	1930–1939	13.99	−0.004 (0.014)	0.051 (0.014)	0.012 (0.014)	5.9 [0.0006]
	1940–1949	14.28	0.005 (0.011)	0.043 (0.011)	−0.003 (0.010)	7.8 [0.0017]
College graduate	1930–1939	0.24	−0.005 (0.002)	0.003 (0.002)	0.002 (0.002)	5.0 [0.0021]
	1940–1949	0.30	−0.003 (0.002)	0.004 (0.002)	0.000 (0.002)	5.0 [0.0018]

Figure 72 shows the first stage from a regression of the following form:

$$S_i = X\pi_{10} + Z_1\pi_{11} + Z_2\pi_{12} + Z_3\pi_{13} + \eta_1$$

where Z_i is the dummy for the first three quarters, and π_i is the coefficient on each dummy. Now we look at what they produced in Figure 72. Consistent with our intuition, the coefficients are all *negative* and significant for the total years of education and the high school graduate dependent variables. Notice, too, that the relationship gets much weaker once we move beyond the groups bound by compulsory schooling: the number of years of schooling for high school students (no effect), and probability of being a college graduate (no effect).

Regarding those college non-results. Ask yourself this question: why should we expect quarter of birth to affect the probability of being a high school graduate, but not on being a college grad? What if we had found quarter of birth predicted high school completion, college completion, post-graduate completion, and total years of schooling beyond high school? Wouldn't it start to seem like this compulsory schooling instrument was not what we thought it was? After all, this quarter of birth instrument really should only impact *high school* completion; since it doesn't bind anyone beyond high school, it shouldn't affect the number of years beyond high school or college completion probabilities. If it did, we might be skeptical of the whole design. But here it didn't, which to me makes it even more convincing that they're identifying a compulsory high school schooling effect.¹¹⁶

Now we look at the second stage for both OLS and 2SLS (which they label TSLS, but means the same thing). Figure 73 shows these results. The authors didn't report the first stage in this table because they reported it in the earlier table we just reviewed.¹¹⁷ For small values, the log approximates a percentage change, so they are finding a 7.1% return for every additional year of schooling, but with 2SLS it's higher (8.9%). That's interesting, because if it was merely ability bias, then we'd expect the OLS estimate to be *too large*, not too small. So something other than mere ability bias must be going on here.

For whatever it's worth, I am personally convinced at this point that quarter of birth is a valid instrument, and that they've identified a causal effect of schooling on earnings, but Angrist and Krueger [1991] want to go further, probably because they want more precision in their estimate. And to get more precision, they load up the first stage with even more instruments. Specifically, they use specifications with 30 dummies (quarter of birth \times year) and 150 dummies (quarter

¹¹⁶ These kinds of falsifications are extremely common in contemporary applied work. This is because many of the identifying assumptions in any research design are simply untestable. And so the burden of proof is on researchers to convince the reader, oftentimes with intuitive and transparent falsification tests.

¹¹⁷ My personal preference is to report everything in the same table, mainly for design reasons. I like fewer tables with each table having more information. In other words, I want someone to look at an instrumental variables table and immediately see the OLS result, the 2SLS result, the first stage relationship, and the *F* statistic on that first stage. See Figure 68 for an example.

Independent variable	(1) OLS	(2) TSLS
Years of education	0.0711 (0.0003)	0.0891 (0.0161)
Race (1 = black)	—	—
SMSA (1 = center city)	—	—
Married (1 = married)	—	—
9 Year-of-birth dummies	Yes	Yes
8 Region-of-residence dummies	No	No
Age	—	—
Age-squared	—	—
χ^2 [dof]	—	25.4 [29]

Figure 73: Angrist and Krueger [1991] OLS and 2SLS results for the effect of education on log weekly earnings.

of birth \times state) as instruments. The idea is that the quarter of birth effect may differ by state and cohort. Because they have more variation in the instrument, the predicted values of schooling also have more variation, which brings down the standard errors.

But at what cost? Many of these instruments are only now weakly correlated with schooling - in some locations, they have almost no correlation, and for some cohorts as well. We got a flavor of that, in fact, in Figure 70 where the later cohorts show less variation in schooling by quarter of birth than the earlier cohorts. What is the effect, then, of reducing the variance in the estimator by loading up the first stage with a bunch of noise?

Work on this starts with Bound et al. [1995] and is often called the “weak instrument” literature. It’s in this paper that we learn some basic practices for determining if we have a weak instrument problem, as well as an understanding of the nature of the bias of IV in finite samples and under different violations of the IV assumptions. Bound et al. [1995] sought to understand what IV was identifying when the first stage was weak, as it was when Angrist and Krueger [1991] loaded up their first stage with 180 instruments, many of which were very weak.

Let’s review Bound et al. [1995] now and consider their model with a single endogenous regressor and a simple constant treatment

effect. The causal model of interest here is as before:

$$y = \beta s + \varepsilon$$

where y is some outcome and s is some endogenous regressor, such as schooling. The matrix of IVs is Z with the first stage equation

$$s = Z'\pi + \eta$$

If ε and η are correlated, then estimating the first equation by OLS would lead to biased results, wherein the OLS bias is:

$$E[\hat{\beta}_{OLS} - \beta] = \frac{C(\varepsilon, s)}{V(s)}$$

We will rename this ratio as $\frac{\sigma_{\varepsilon\eta}}{\sigma_s^2}$. It can be shown that the bias of 2SLS is approximately:

$$E[\hat{\beta}_{2SLS} - \beta] \approx \frac{\sigma_{\varepsilon\eta}}{\sigma_\eta^2} \frac{1}{F+1}$$

where F is the population analogy of the F -statistic for the joint significance of the instruments in the first stage regression. If the first stage is weak, then $F \rightarrow 0$, then the bias of 2SLS approaches $\frac{\sigma_{\varepsilon\eta}}{\sigma_\eta^2}$. But if the first stage is very strong, $F \rightarrow \infty$, then the 2SLS bias goes to 0.

Returning to our rhetorical question from earlier, what was the cost of adding instruments without predictive power? Adding more weak instruments causes the first stage F statistic to approach zero and increase the bias of 2SLS.

What if the model is “just identified”, meaning there’s the same number of instruments as there are endogenous variables? Bound et al. [1995] studied this empirically, replicating Angrist and Krueger [1991], and using simulations. Figure 74 shows what happens once they start adding in controls. Notice that as they do, the F statistic on the excludability of the instruments falls from 13.5 to 4.7 to 1.6. So by the F statistic, they are already running into a weak instrument once they include the 30 quarter of birth \times year dummies, and I think that’s because as we saw, the relationship between quarter of birth and schooling got smaller for the later cohorts.

Next, they added in the weak instruments – all 180 of them – which is shown in Figure 75. And here we see that the problem persists. The instruments are weak, and therefore the bias of the 2SLS coefficient is close to that of the OLS bias.

But the really damning part of the Bound et al. [1995] paper was their simulation. The authors write:

“To illustrate that second-stage results do not give us any indication of the existence of quantitatively important finite-sample biases, we reestimated Table 1, columns (4) and (6) and Table 2, columns (2)

Table 1. Estimated Effect of Completed Years of Education on Men's Log Weekly Earnings
(standard errors of coefficients in parentheses)

	(1) OLS	(2) IV	(3) OLS	(4) IV	(5) OLS	(6) IV
Coefficient	.063 (.000)	.142 (.033)	.063 (.000)	.081 (.016)	.063 (.000)	.060 (.029)
F (excluded instruments)		13.486		4.747		1.613
Partial R ² (excluded instruments, $\times 100$)		.012		.043		.014
F (overidentification)		.932		.775		.725
<i>Age Control Variables</i>						
Age, Age ²	x	x			x	x
9 Year of birth dummies			x	x	x	x
<i>Excluded Instruments</i>						
Quarter of birth		x		x		x
Quarter of birth \times year of birth				x		x
Number of excluded instruments	3		30		28	

NOTE: Calculated from the 5% Public-Use Sample of the 1980 U.S. Census for men born 1930–1939. Sample size is 329,509. All specifications include Race (1 = black), SMSA (1 = central city), Married (1 = married, living with spouse), and 8 Regional dummies as control variables. F (first stage) and partial

Figure 74: Bound et al. [1995] OLS and 2SLS results for the effect of education on log weekly earnings.

and (4), using **randomly** generated information in place of the actual quarter of birth, following a suggestion by Alan Krueger. The means of the estimated standard errors reporting in the last row are quite close to the actual standard deviations of the 500 estimates for each model. ... It is striking that the second-stage results reported in Table 3 look quite reasonable even with no information about educational attainment in the simulated instruments. They give no indication that the instruments were randomly generated. ... On the other hand, the F statistics on the excluded instruments in the first-stage regressions are always near their expected value of essentially 1 and do give a clear indication that the estimates of the second-stage coefficients suffer from finite-sample biases.”

So, what can you do if you have weak instruments. First, you can use a just identified model with your strongest IV. Second, you can use a limited information maximum likelihood estimator (LIML). This is approximately median unbiased for over identified constant effects models. It provides the same asymptotic distribution as 2SLS under homogenous treatment effects, but provides a finite-sample bias reduction.

But, let’s be real for a second. If you have a weak instrument problem, then you only get so far by using LIML or estimating a just identified model. The real solution for a weak instrument problem is *get better instruments*. Under homogenous treatment effects, you’re always identifying the same effect so there’s no worry about a complier only parameter. So you should just continue searching for stronger instruments that simultaneously satisfy the exclusion restriction.¹¹⁸

In conclusion, circling back to where we started, I think we’ve learned a lot about instrumental variables and why it is so powerful. The estimators based on this design are capable of identifying causal effects when your data suffer from selection on unobservables. Since selection on unobservables is believed to be very common, this is a

¹¹⁸ Good luck with that. Seriously, good luck.

Table 2. Estimated Effect of Completed Years of Education on Men's Log Weekly Earnings, Controlling for State of Birth (standard errors of coefficients in parentheses)

	(1) OLS	(2) IV	(3) OLS	(4) IV
Coefficient	.063 (.000)	.083 (.009)	.063 (.000)	.081 (.011)
<i>F</i> (excluded instruments)		2.428		1.869
Partial <i>R</i> ² (excluded instruments, $\times 100$)		.133		.101
<i>F</i> (overidentification)		.919		.917
<i>Age Control Variables</i>				
Age, Age ²			x	x
9 Year of birth dummies	x	x	x	x
<i>Excluded Instruments</i>				
Quarter of birth		x		x
Quarter of birth \times year of birth		x		x
Quarter of birth \times state of birth		x		x
Number of excluded instruments	180		178	

NOTE: Calculated from the 5% Public-Use Sample of the 1980 U.S. Census for men born 1930–1939. Sample size is 329,509. All specifications include Race (1 = black), SMSA (1 = central city), Married (1 = married, living with spouse), 8 Regional dummies, and 50 State of Birth dummies as control variables. *F* (first stage) and partial *R*² are for the instruments in the first stage of IV estimation. *F* (overidentification) is that suggested by Basmann (1960).

very useful methodology for addressing it. But, that said, we also have learned some of its weaknesses, and hence why some people eschew it. Let's now move to heterogeneous treatment effects so that we can better understand some of its limitations a bit better.

Heterogenous treatment effects

Now we turn to the more contemporary pedagogy where we relax the assumption that treatment effects are the same for every unit. Now we will allow for each unit to have a unique response to the treatment, or

$$Y_i^1 - Y_i^0 = \delta_i$$

Note that the treatment effect parameter now differs by individual i .

The main questions we have now are: (1) what is IV estimating when we have heterogeneous treatment effects, and (2) under what assumptions will IV identify a causal effect with heterogeneous treatment effects? The reason why this matters is that once we introduce

Figure 75: Bound et al. [1995] OLS and 2SLS results for the effect of education on log weekly earnings with the 100+ weak instruments.

heterogenous treatment effects, we introduce a distinction between the internal validity of a study and its external validity. Internal validity means our strategy identified a causal effect *for the population we studied*. But external validity means the study's finding applied to *different populations* (not in the study). As we'll see, under homogenous treatment effects, there is no such tension between external and internal validity because everyone has the same treatment effect. But under heterogenous treatment effects, there is a huge tension; the tension is so great, in fact, that it may even undermine an otherwise valid IV design.

Heterogenous treatment effects are built on top of the potential outcomes notation, with a few modifications. Since now we have two arguments - D and Z - we have to modify the notation slightly. We say that Y is a function of D and Z as $Y_i(D_i = 0, Z_i = 1)$, which is represented as $Y_i(0, 1)$.

Potential *outcomes* as we have been using the term refers to the Y variable, but now we have a new potential variable – potential *treatment status* (as opposed to observed treatment status). Here's the characteristics:

- $D_i^1 = i$'s treatment status when $Z_i = 1$
- $D_i^0 = i$'s treatment status when $Z_i = 0$
- And observed treatment status is based on a treatment status switching equations:

$$\begin{aligned} D_i &= D_i^0 + (D_i^1 - D_i^0)Z_i \\ &= \pi_0 + \pi_1 Z_i + \phi_i \end{aligned}$$

where $\pi_{0i} = E[D_i^0]$, $\pi_{1i} = (D_i^1 - D_i^0)$ is the heterogenous causal effect of the IV on D_i , and $E[\pi_{1i}]$ = the average causal effect of Z_i on D_i .

There are considerably more assumptions necessary for identification once we introduce heterogenous treatment effects – specifically five assumptions. We now review each of them. And to be concrete, I will use repeatedly as an example the effect of military service on earnings using a draft lottery as the instrumental variable [Angrist, 1990].

First, as before, there is a stable unit treatment value assumption (SUTVA) which states that the potential outcomes for each person i are unrelated to the treatment status of other individuals. The assumption states that if $Z_i = Z'_i$, then $D_i(Z) = D_i(Z')$. And if $Z_i = Z'_i$ and $D_i = D'_i$, then $Y_i(D, Z) = Y_i(D', Z')$. An violation of SUTVA would be if the status of a person at risk of being drafted was affected by the draft status of others at risk of being drafted. Such spillovers violate SUTVA.¹¹⁹

¹¹⁹ Probably no other identifying assumption is given shorter shrift than SUTVA. Rarely is it mentioned in applied studies, let alone taken seriously.

Second, there is the independence assumption. The independence assumption is also sometimes called the “as good as random assignment” assumption. It states that the IV is independent of the potential outcomes and potential treatment assignments. Notationally, it is

$$\{Y_i(D_i^1, 1), Y_i(D_i^0, 0), D_i^1, D_i^0\} \perp\!\!\!\perp Z_i$$

The independence assumption is sufficient for a causal interpretation of the reduced form:

$$\begin{aligned} E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0] &= E[Y_i(D_i^1, 1)|Z_i = 1] - E[Y_i(D_i^0, 0)|Z_i = 0] \\ &= E[Y_i(D_i^1, 1)] - E[Y_i(D_i^0, 0)] \end{aligned}$$

Independence means that the first stage measures the causal effect of Z_i on D_i :

$$\begin{aligned} E[D_i|Z_i = 1] - E[D_i|Z_i = 0] &= E[D_i^1|Z_i = 1] - E[D_i^0|Z_i = 0] \\ &= E[D_i^1 - D_i^0] \end{aligned}$$

An example of this is if Vietnam conscription for military service was based on randomly generated draft lottery numbers. The assignment of draft lottery number was independent of potential earnings or potential military service because it was “as good as random”.

Third, there is the exclusion restriction. The exclusion restriction states that any effect of Z on Y must be via the effect of Z on D . In other words, $Y_i(D_i, Z_i)$ is a function of D_i only. Or formally:

$$Y_i(D_i, 0) = Y_i(D_i, 1) \text{ for } D = 0, 1$$

Again, our Vietnam example. In the Vietnam draft lottery, an individual's earnings potential as a veteran or a non-veteran are assumed to be the same regardless of draft eligibility status. The exclusion restriction would be violated if low lottery numbers affected schooling by people avoiding the draft. If this was the case, then the lottery number would be correlated with earnings for at least two cases. One, through the instrument's effect on military service. And two, through the instrument's effect on schooling. The implication of the exclusion restriction is that a random lottery number (independence) does not therefore imply that the exclusion restriction is satisfied. These are different assumptions.

Fourth is the first stage. IV under heterogeneous treatment effects requires that Z be correlated with the endogenous variable such that

$$E[D_i^1 - D_i^0] \neq 0$$

Z has to have some statistically significant effect on the average probability of treatment. An example would be having a low lottery

number. Does it increase the average probability of military service? If so, then it satisfies the first stage requirement. Note, unlike independence and exclusion, the first stage is testable as it is based solely on D and Z , both of which you have data on.

And finally, the monotonicity assumption. This is only strange at first glance, but is actually quite intuitive. Monotonicity requires that the instrumental variable (weakly) operate in the same direction on all individual units. In other words, while the instrument may have no effect on some people, all those who are affected are affected in the same direction (i.e., positively or negative, but not both). We write it out like this:

$$\text{Either } \pi_{1i} \geq 0 \text{ for all } i \text{ or } \pi_{1i} \leq 0 \text{ for all } i = 1, \dots, N$$

What this means, as an example, using our military draft example, is that draft eligibility may have no effect on the probability of military service for some people, like patriots, but when it does have an effect, it shifts them all into service, or out of service, but not both. The reason that we have to make this assumption is that without monotonicity, IV estimators are not guaranteed to estimate a weighted average of the underlying causal effects of the affected group.

If all five assumptions are satisfied, then we have a valid IV strategy. But that being said, while valid, it is not doing what it was doing when we had homogenous treatment effects. What, then, is the IV strategy estimating under heterogenous treatment effects? Answer: the local average treatment effect (LATE) of D on Y :

$$\begin{aligned}\delta_{IV,LATE} &= \frac{\text{Effect of } Z \text{ on } Y}{\text{Effect of } Z \text{ on } D} \\ &= \frac{E[Y_i(D_i^1, 1) - Y_i(D_i^0, 0)]}{E[D_i^1 - D_i^0]} \\ &= E[(Y_i^1 - Y_i^0) | D_i^1 - D_i^0 = 1]\end{aligned}$$

The LATE parameters is the average causal effect of D on Y for those whose treatment status was changed by the instrument, Z . For instance, IV estimates the average effect of military service on earnings for the subpopulations who enrolled in military service *because of the draft* but who would not have served otherwise. It doesn't identify the causal effect on patriots who always serve, for instance, because those individuals did not have their military service pushed or pulled by the draft number. It also won't tell us the effect of military service on those who were exempted from military service for medical reasons.¹²⁰

The LATE framework has even more jargon, so let's review it now. The LATE framework partitions the population of units with

¹²⁰ We have reviewed the properties of IV with heterogenous treatment effects using a very simple dummy endogenous variable, dummy IV, and no additional controls example. The intuition of LATE generalizes to most cases where we have continuous endogenous variables and instruments, and additional control variables, as well.

an instrument into potentially four mutually exclusive groups. Those groups are:

1. Compliers: this is the subpopulation whose treatment status is affected by the instrument in the correct direction. That is, $D_i^1 = 1$ and $D_i^0 = 0$.
2. Defiers: this is the subpopulation whose treatment status is affected by the instrument in the wrong direction. That is, $D_i^1 = 0$ and $D_i^0 = 1$.¹²¹
3. Never takers: this is the subpopulation of units that never take the treatment regardless of the value of the instrument. So, $D_i^1 = D_i^0 = 0$. They simply never take the treatment.¹²²
4. Always takers: this is the subpopulation of units that always take the treatment regardless of the value of the instrument. So, $D_i^1 = D_i^0 = 1$. They simply always take the instrument.¹²³

As outlined above, with all five assumptions satisfied, IV estimates the average treatment effect for compliers. Contrast this with the traditional IV pedagogy with homogenous treatment effects. In that situation, compliers have the same treatment effects as non-compliers, so the distinction is irrelevant. Without further assumptions, LATE is not informative about effects on never-takers or always-takers because the instrument does not affect their treatment status.

Does this matter? Yes, absolutely. It matters because in most applications, we would be mostly interested in estimating the average treatment effect on the whole population, but that's not usually possible with IV.¹²⁴

Now that we have reviewed the basic idea and mechanics of instrumental variables, including some of the more important tests associated with it, let's get our hands dirty with some data. We'll work with a couple of datasets now to help you better understand how to implement 2SLS in real data.

Stata exercise #1: College in the county

We will once again look at the returns to schooling since it is such a historically popular topic for causal questions in labor. In this application, we will simply show how to use the Stata command `ivregress` with 2SLS, calculate the first stage F statistic, and compare the 2SLS results with the OLS results. I will be keeping it simple, because my goal is just to help the reader become familiarized with the procedure.

The data comes from the NLS Young Men Cohort of the National Longitudinal Survey. This data began in 1966 with 5,525 men aged

¹²¹ So for instance, say that we have some instrument for attending a private school. Compliers go to the school if they win the lottery, and don't go to the school if they don't. Defiers attend the school if they don't win, but attend the school if they do win. Defiers sound like jerks.

¹²² Sticking with our private school lottery example. This is a group of people who believe in public education, and so even if they win the lottery, they won't go. They're never-takers; they never go to private school no matter what.

¹²³ This is a group of people who always send their kids to private school, regardless of the number on their voucher lottery.

¹²⁴ This identification of the LATE under heterogenous treatment effects material was worked out in Angrist et al. [1996]. See it for more details.

14-24 and continued to follow up with them through 1981. These data come from 1966, the baseline survey, and there's a number of questions related to local labor markets. One of them is whether the respondent lives in the same county as a 4-year (and a 2-year) college.

Card [1995] is interested in estimating the following regression equation:

$$Y_i = \alpha + \delta S_i + \gamma X_i + \varepsilon_i$$

where Y is log earnings, S is years of schooling, X is a matrix of exogenous covariates and ε is an error term that contains among other things unobserved ability. Under the assumption that ε contains ability, and ability is correlated with schooling, then $C(S, \varepsilon) \neq 0$ and therefore schooling is biased. Card [1995] proposes therefore an instrumental variables strategy whereby he will instrument for schooling with the college-in-the-county dummy variable.

It is worth asking ourselves why the presence of a four year college in one's county would increase schooling. The main reason that I can think of is that the presence of the 4-year-college increases the likelihood of going to college by lowering the costs, since the student can live at home. This therefore means, though, that we are selecting on a group of compliers whose behavior is affected by the variable. Some kids, in other words, will always go to college regardless of whether a college is in their county, and some will never go despite the presence of the nearby college. But there may exist a group of compliers who go to college only because their county has a college, and if I'm right that this is primarily picking up people going because they can attend while living at home, then it's necessarily people at some margin who attend only because college became slightly cheaper. This is, in other words, a group of people who are liquidity constrained. And if we believe the returns to schooling for this group is different than that of the always-takers, then our estimates may not represent the ATE. Rather, they would represent the LATE. But in this case, that might actually be an interesting parameter since it gets at the issue of lowering costs of attendance for poorer families.

Here we will do some simple analysis based on Card [1995].

- . scuse card
- . reg lwage educ exper black south married smsa
- . ivregress 2sls lwage (educ=nearc4) exper black south married smsa, first
- . reg educ nearc4 exper black south married smsa
- . test nearc4

And our results from this analysis have been arranged into Table 28. First, we report our OLS results. For every one year additional

of schooling, respondents' earnings increase by approximately 7.1%. Next we estimated 2SLS using the `ivregress 2sls` command in Stata. Here we find a much larger return to schooling than we had found using OLS - around 75% larger in fact. But let's look at the first stage first. We find that the college in the county is associated with a 0.327 more years of schooling. This is highly significant ($p < 0.001$). The F -statistic exceeds 15, suggesting we don't have a weak instrument problem. The return to schooling associated with this 2SLS estimate is 0.124 – that is, for every additional year of schooling, earnings increases by 12.4%. Other covariates are listed if you're interested in studying them as well.

Dependent variable	Log earnings	
	OLS	2SLS
educ	0.071*** (0.003)	0.124** (0.050)
exper	0.034*** (0.002)	0.056*** (0.020)
black	-0.166*** (0.018)	-0.116** (0.051)
south	-0.132*** (0.015)	-0.113*** (0.023)
married	-0.036*** (0.003)	-0.032*** (0.005)
smsa	0.176*** (0.015)	0.148*** (0.031)

First Stage Instrument		
College in the county		0.327*** (0.082)
Robust standard error		
F statistic for IV in first stage		15.767
N	3,003	3,003
Mean Dependent Variable	6.262	6.262
Std. Dev. Dependent Variable	0.444	0.444

Standard errors in parenthesis. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 28: OLS and 2SLS regressions of Log Earnings on Schooling

Why would the return to schooling be so much larger for the compliers than for the general population? After all, we showed earlier that if this was simply ability bias, then we'd expect the 2SLS coefficient to be *smaller* than the OLS coefficient, because ability bias implies that the coefficient on schooling is *too large*. Yet we're finding the opposite. So a couple of things it could be. First, it could be that schooling has measurement error. Measurement error would bias the coefficient towards zero, and 2SLS would recover its true value. But

I find this explanation to be unlikely, because I don't foresee people really not knowing with accuracy how many years of schooling they currently have. Which leads us to the other explanation, and that is that compliers have larger returns to schooling. But why would this be the case? Assuming that the exclusion restriction holds, then why would compliers returns be so much larger? We've already established that these people are likely being shifted into more schooling because they live with their parents, which suggests that the college is lowering the marginal cost of going to college. All we are left saying is that for some reason, the higher marginal cost of attending college is causing these people to under invest in schooling; that in fact their returns are much higher. I welcome your thoughts, though, on why this number might be so different.

Stata exercise #2: Fulton fish markets

The second exercise that we'll be doing is based on [Graddy \[2006\]](#). My understanding is that Graddy hand collected these data herself by recording prices of fish at the actual Fulton fish market. I'm not sure if that is true, but I like to believe it's true, because I like to believe in shoe leather research of that kind. Anyhow, the Fulton Fish Market operated in NYC on Fulton Street for 150 years. In November 2005, they moved it from the lower Manhattan to a large facility building for the market in the South Bronx. At the time of the article's writing, it was called the New Fulton Fish Market. It's one of the world's largest fish markets, second only to the Tsukiji in Tokyo.

This is an interesting market because fish are heterogenous, highly differentiated products. There are anywhere between 100 to 300 different varieties of fish sold at the market. There are over 15 different varieties of shrimp alone. Within each variety, there's small fish, large fish, medium fish, fish just caught, fish that have been around a while. There's so much heterogeneity in fact that customers often want to examine fish personally. You get the picture. This fish market functions just like a two-sided platform matching buyers to sellers, which is made more efficient by the thickness the market produces. It's not surprising, therefore, that Graddy found the market such an interesting thing to study.

Let's move to the data. I want us to estimate the price elasticity of demand for fish, which makes this problem much like the problem that Philip Wright faced in that price and quantity are determined simultaneously. The elasticity of demand is a sequence of quantity and price pairs, but with only one pair observed at a given point in time. In that sense, the demand curve is itself a sequence of potential outcomes (quantity) associated with different potential treatments

(price). This means the demand curve is itself a real object, but mostly unobserved. Therefore, to trace out the elasticity, we need an instrument that is correlated with supply only. Graddy proposes a few of them, all of which have to do with the weather at sea in the days before the fish arrived to market.

The first instrument is the average max last 2 days wave height. The model we are interested in estimating is:

$$Q = \alpha + \delta P + \gamma X + \varepsilon$$

where Q is log quantity of whiting sold in pounds, P is log average daily price per pound, X are day of the week dummies and a time trend, and ε is the structural error term. Table 29 presents the results from estimating this equation with OLS (first column) and 2SLS (second column). The OLS estimate of the elasticity of demand is -0.549. It could've been anything given price is determined by how many sellers and how many buyers there are at the Market on any given day. But when we use the average wave height as the instrument for price, we get a -0.96 price elasticity of demand. A 10% increase in the price causes quantity to decrease by 9.6%. The instrument is strong ($F > 22$). For every one unit increase in the wave height, price rose 10%.

I suppose the question we have to ask ourselves, though, is what exactly is this instrument doing to supply. What are higher waves doing exactly? It's making it more difficult to fish, but is it also changing the composition of the fish caught? If so, then it would seem that the exclusion restriction is violated because that would mean the wave height is directly causing fish composition to change which will directly determine quantities bought and sold.

Now let's look at a different instrument: windspeed. Specifically, it's the 3 day lagged max windspeed. We present these results in Table 29. Here we see something we did not see before, which is that this is a weak instrument. The F statistic is less than 10 (approximately 6.5). And correspondingly, the estimated elasticity is twice as large as what we found with wave height. Thus we know from our earlier discussion of weak instruments that this estimate is likely biased, and therefore less reliable than the previous one – even though the previous one itself (1) may not convincingly satisfy the exclusion restriction and (2) is at best a LATE relevant to compliers only. But as we've said, if we think that the compliers' causal effects are similar to that of the broader population, then the LATE may itself be informative and useful.

We've reviewed the use of IV in identifying causal effects when some regressor is endogenous in observational data. But increasingly, you're seeing it used with randomized trials. In many randomized

Dependent variable	Log quantity	
	OLS	2SLS
Log(Price)	-0.549*** (0.184)	-0.960** (0.406)
Monday	-0.318 (0.227)	-0.322 (0.225)
Tuesday	-0.684*** (0.224)	-0.687*** (0.221)
Wednesday	-0.535** (0.221)	-0.520** (0.219)
Thursday	0.068 (0.221)	0.106 (0.222)
Time trend	-0.001 (0.003)	-0.003 (0.003)

First Stage Instrument		
Average wave height		0.103***
Robust standard error		(0.022)
F statistic for IV in first stage		22.638
N	97	97
Mean Dependent Variable	8.086	8.086
Std. Dev. Dependent Variable	0.765	0.765

Table 29: OLS and 2SLS regressions of Log Quantity on Log Price with wave height instrument

Standard errors in parenthesis. * p<0.10, ** p<0.05, *** p<0.01

Dependent variable	Log quantity	
	OLS	2SLS
Log Price	-0.549*** (0.184)	-1.960** (0.873)
Monday	-0.318 (0.227)	-0.332 (0.281)
Tuesday	-0.684*** (0.224)	-0.696** (0.277)
Wednesday	-0.535** (0.221)	-0.482* (0.275)
Thursday	0.068 (0.221)	0.196 (0.285)
Time trend	-0.001 (0.003)	-0.007 (0.005)

First Stage Instrument		
Wind Speed		0.017**
Robust standard error		(0.007)
F statistic for IV in first stage		6.581
N	97	97
Mean Dependent Variable	8.086	8.086
Std. Dev. Dependent Variable	0.765	0.765

Table 30: OLS and 2SLS regressions of Log Quantity on Log Price with windspeed instrument

Standard errors in parenthesis. * p<0.10, ** p<0.05, *** p<0.01

trials, participation is voluntary among those randomly chosen to be in the treatment group. On the other hand, persons in the control group usually don't have access to the treatment. Only those who are particularly likely to benefit from treatment therefore will probably take up treatment which almost always leads to positive selection bias. If you just compare means between treated and untreated individuals using OLS, you will obtain biased treatment effects even for the randomized trial due to non-compliance. So a solution is to instrument for treatment with whether you were offered treatment and estimate the LATE. Thus even when treatment itself is randomly assigned, it is common for people to use a randomized lottery as an instrument for participation. For a modern example of this, see Baicker et al. [2013] who used the randomized lottery to be on Oregon's Medicaid as an instrument for being on Medicaid.

In conclusion, instrumental variables is a powerful design for identifying causal effects when your data suffer from selection on unobservables. But even with that in mind, it has many limitations that has in the contemporary period caused many applied researchers to eschew it. First, it only identifies the LATE under heterogeneous treatment effects, and that may or may not be a policy relevant variable. Its value ultimately depends on how closely the compliers' average treatment effect resembles that of the other subpopulations'. Second, unlike RDD which has only 1 main identifying assumption (the continuity assumption), IV has up to 5 assumptions! Thus, you can immediately see why people find IV estimation less credible – not because it fails to identify a causal effect, but rather because it's harder and harder to imagine a pure instrument that satisfies all five conditions. But all this is to say, IV is an important strategy and sometimes the opportunity to use it will come along, and you should be prepared for when that happens by understanding it and how to implement it in practice.

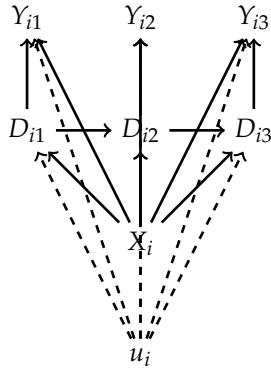
Panel data

Introduction

One of the most important tools in the causal inference toolkit are the panel data estimators. These are estimators designed explicitly for longitudinal data – the repeated observing of a unit over time. Under certain situations, repeatedly observing the same unit over time can overcome a particular kind of omitted variable bias, though not all kinds. While it is possible that observing the same unit over time will not resolve the bias, there are still many applications where it can, and that's why this method is so important. We review first the DAG describing just such a situation, followed by discussion of a paper, and then present a dataset exercise in Stata.

DAG Example

Before I dig into the technical assumptions and estimation methodology for panel data techniques, I wanted to review a simple DAG illustrating those assumptions. This DAG comes from [Imai and Kim \[2017\]](#). Let's say that we have data on a column of outcomes, Y_i , which appear in three time periods. In other words, Y_{i1} , Y_{i2} , and Y_{i3} where i indexes a particular unit and $t = 1, 2, 3$ index the time period where each i unit is observed. Likewise, we have a matrix of covariates, D_i , which also vary over time – D_{i1} , D_{i2} , and D_{i3} . And finally there exists a single unit-specific unobserved variable, u_i , which varies across units, but which does not vary over time for that unit. Hence the reason that there is no $t = 1, 2, 3$ subscript for our u_i variable. Key to this variable is (a) it is unobserved in the dataset, (b) it is unit-specific, and (c) it does not change over time for a given unit i . Finally there exists some unit-specific time-invariant variable, X_i . Notice that it doesn't change over time, just u_i , but unlike u_i it is observed.



As this is the busiest DAG we've seen so far, it merits some discussion. First, note that D_{it} causes both its own outcome Y_{it} is also correlated with the next period D_{it+1} . Secondly, u_i is correlated with all the Y_{it} and D_{it} variables, which technically makes D_{it} endogenous since u_i is unobserved and therefore gets absorbed into a composite error term. Thirdly, there is *no* time-varying unobserved confounder correlated with D_{it} - the only confounder is u_i , which we call the unobserved heterogeneity. Fourth, past outcomes do not directly affect current outcomes (i.e., no direct edge between the Y_{it} variables). Fifth, past outcomes do not directly affect current treatments (i.e., no direct edge from Y_{it-1} to D_{it}). And finally, past treatments, D_{it-1} do not directly affect current outcomes, Y_{it} (i.e., no direct edge from D_{it-1} and Y_{it}). It is under these assumptions that we can use a particular panel method called *fixed effects* to isolate the causal effect of D on Y .

What might an example of this be? Let's return to our story about the returns to education. Let's say that we are interested in the effect of schooling on earnings, and schooling is partly determined by unchanging genetic factors which themselves determine unobserved ability, like intelligence, contentiousness and motivation [Conley and Fletcher, 2017]. If we observe the same people's time varying earnings and schoolings over time, then if the situation described by the above DAG describes both the directed edges and *the missing edges*, then we can use panel fixed effects models to identify the causal effect of schooling on earnings.

Estimation

When we use the term "panel data", what do we mean? We mean a dataset where we observe the same units (individuals, firms, countries, schools, etc.) over more than one time period. Often our outcome variable depends on several factors, some of which are observed and some of which are unobserved in our data, and insofar as the unobserved variables are correlated with the treatment variable,

then the treatment variable is endogenous and correlations are not estimates of a causal effect. This chapter focuses on the conditions under which a correlation between D and Y reflects a causal effect even with unobserved variables that are correlated with the treatment variable. Specifically, if these omitted variables are *constant* over time, then even if they are heterogeneous across units, we can use panel data estimators to consistently estimate the effect of our treatment variable on outcomes.

There are several different kinds of estimators for panel data, but we will in this chapter only cover two: pooled ordinary least squares (POLS) and fixed effects (FE).¹²⁵

First we need to set up our notation. With some exceptions, panel methods are usually based on the traditional notation and not the potential outcomes notation. One exception, though, includes the matrix completion methods by [Athey et al. \[2017\]](#), but at the moment, that material is not included in this version. So we will use, instead, the traditional notation for our motivation.

Let Y and $D \equiv (D_1, D_2, \dots, D_k)$ be observable random variables and u be an unobservable random variable. We are interested in the partial effects of variable D_j in the population regression function:

$$E[Y|D_1, D_2, \dots, D_k, u]$$

We observe a sample of $i = 1, 2, \dots, N$ cross-sectional units for $t = 1, 2, \dots, T$ time periods (a balanced panel). For each unit i , we denote the observable variables for all time periods as $\{(Y_{it}, D_{it}) : t = 1, 2, \dots, T\}$.¹²⁶ Let $D_{it} \equiv (D_{it1}, D_{it2}, \dots, D_{itk})$ is a $1 \times K$ vector. We typically assume that the actual cross-sectional units (e.g., individuals in a panel) are identical and independent draws from the population in which case $\{Y_i, D_i, u_i\}_{i=1}^N \sim i.i.d.$, or cross-sectional independence. We describe the main observables, then, as $Y_i \equiv (Y_{i1}, Y_{i2}, \dots, Y_{iT})'$ and $D_i \equiv (D_{i1}, D_{i2}, \dots, D_{iT})$.

It's helpful now to illustrate the actual stacking of individual units across their time periods. A single unit i will have multiple time periods t

$$Y_i = \begin{pmatrix} Y_{i1} \\ \vdots \\ Y_{it} \\ \vdots \\ Y_{iT} \end{pmatrix}_{T \times 1} \quad D_i = \begin{pmatrix} D_{i,1,1} & D_{i,1,2} & D_{i,1,j} & \dots & D_{i,1,K} \\ \vdots & \vdots & \vdots & & \vdots \\ D_{i,t,1} & D_{i,t,2} & D_{i,t,j} & \dots & D_{i,t,K} \\ \vdots & \vdots & \vdots & & \vdots \\ D_{i,T,1} & D_{i,T,2} & D_{i,T,j} & \dots & D_{i,T,K} \end{pmatrix}_{T \times K}$$

And the entire panel itself with all units included will look like this:

¹²⁵ A common third type of panel estimator is the random effects estimator, but in my experience, I have used it less often than fixed effects, so I decided to omit it. Again, this is not because it is unimportant. It is important. I just have chosen to do fewer things in more detail based on whether I think they qualify as the most common methods used in the present period by applied empiricists. See [Wooldridge \[2010\]](#) for a more comprehensive treatment, though, of all panel methods including random effects.

¹²⁶ For simplicity, I'm ignoring the time-invariant observations, X_i from our DAG for reasons that will hopefully soon be made clear.

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_i \\ \vdots \\ Y_N \end{pmatrix}_{NT \times 1} \quad D = \begin{pmatrix} D_1 \\ \vdots \\ D_i \\ \vdots \\ D_N \end{pmatrix}_{NT \times K}$$

For a randomly drawn cross-sectional unit i , the model is given by

$$Y_{it} = \delta D_{it} + u_i + \varepsilon_{it}, \quad t = 1, 2, \dots, T$$

As always, we use our schooling-earnings example for motivation. Let Y_{it} be log earnings for a person i in year t . Let D_{it} be schooling for person i in year t . Let δ be the returns to schooling. Let u_i be the sum of all time-invariant person-specific characteristics, such as unobserved ability. This is often called the *unobserved heterogeneity*. And let ε_{it} be the time-varying unobserved factors that determine a person's wage in a given period. This is often called the idiosyncratic error. We want to know what happens when we regress Y_{it} on D_{it} .

Pooled OLS The first estimator we will discuss is the pooled Ordinary Least Squares or POLS estimator. When we ignore the panel structure and regress Y_{it} on D_{it} we get

$$Y_{it} = \delta D_{it} + \eta_{it}; \quad t = 1, 2, \dots, T$$

with composite error $\eta_{it} \equiv c_i + \varepsilon_{it}$. The main assumption necessary to obtain consistent estimates for δ is:

$$E[\eta_{it}|D_{i1}, D_{i2}, \dots, D_{iT}] = E[\eta_{it}|D_{it}] = 0 \text{ for } t = 1, 2, \dots, T$$

While our DAG did not include ε_{it} , this would be equivalent to assuming that the unobserved heterogeneity, c_i , was uncorrelated with D_{it} for all time periods.

But this is not an appropriate assumption in our case because our DAG explicitly links the unobserved heterogeneity to both the outcome and the treatment in each period. Or using our schooling-earnings example, schooling is likely based on unobserved background factors, u_i , and therefore without controlling for it, we have omitted variable bias and $\hat{\delta}$ is biased. No correlation between D_{it} and η_{it} necessarily means no correlation between the unobserved u_i and D_{it} for all t and that is just probably not a credible assumption. An additional problem is that η_{it} is serially correlated for unit i since u_i is present in each t period. And thus pooled OLS standard errors are also invalid.

Fixed Effects (Within Estimator) Let's rewrite our unobserved effects model so that this is still firmly in our minds:

$$Y_{it} = \delta D_{it} + u_i + \varepsilon_{it}; \quad t = 1, 2, \dots, T$$

If we have data on multiple time periods, we can think of u_i as **fixed effects** to be estimated. OLS estimation with fixed effects yields

$$(\hat{\delta}, \hat{u}_1, \dots, \hat{u}_N) = \underset{b, m_1, \dots, m_N}{\operatorname{argmin}} \sum_{i=1}^N \sum_{t=1}^T (Y_{it} - D_{it}b - m_i)^2$$

this amounts to including N individual dummies in regression of Y_{it} on D_{it} .

The first-order conditions (FOC) for this minimization problem are:

$$\sum_{i=1}^N \sum_{t=1}^T D'_{it} (Y_{it} - D_{it}\hat{\delta} - \hat{u}_i) = 0$$

and

$$\sum_{t=1}^T (Y_{it} - D_{it}\hat{\delta} - \hat{u}_i) = 0$$

for $i = 1, \dots, N$.

Therefore, for $i = 1, \dots, N$,

$$\hat{u}_i = \frac{1}{T} \sum_{t=1}^T (Y_{it} - D_{it}\hat{\delta}) = \bar{Y}_i - \bar{D}_i\hat{\delta},$$

where

$$\bar{D}_i \equiv \frac{1}{T} \sum_{t=1}^T D_{it}; \quad \bar{Y}_i \equiv \frac{1}{T} \sum_{t=1}^T Y_{it}$$

Plug this result into the first FOC to obtain:

$$\begin{aligned} \hat{\delta} &= \left(\sum_{i=1}^N \sum_{t=1}^T (D_{it} - \bar{D}_i)' (D_{it} - \bar{D}_i) \right)^{-1} \left(\sum_{i=1}^N \sum_{t=1}^T (D_{it} - \bar{D}_i)' (\bar{Y}_i - \bar{Y}) \right) \\ \hat{\delta} &= \left(\sum_{i=1}^N \sum_{t=1}^T \ddot{D}'_{it} \ddot{D}_{it} \right)^{-1} \left(\sum_{i=1}^N \sum_{t=1}^T \ddot{D}'_{it} \ddot{D}_{it} \right) \end{aligned}$$

with time-demeaned variables $\ddot{D}_{it} \equiv D_{it} - \bar{D}$, $\ddot{Y}_{it} \equiv Y_{it} - \bar{Y}_i$.

In case it isn't clear, though, running a regression with the time-demeaned variables $\ddot{Y}_{it} \equiv Y_{it} - \bar{Y}_i$ and $\ddot{D}_{it} \equiv D_{it} - \bar{D}$ is *numerically equivalent* to a regression of Y_{it} on D_{it} and unit specific dummy variables. Hence the reason this is sometimes called the "within" estimator, and sometimes called the "fixed effects" estimator. They are the same thing.¹²⁷

Even better, the regression with the time demeaned variables is consistent for δ even when $C[D_{it}, u_i] \neq 0$ because time-demeaning

¹²⁷ One of the things you'll find over time is that things have different names, depending on the author and tradition, and those names are often completely uninformative.

eliminates the unobserved effects. Let's see this now:

$$\begin{aligned} Y_{it} &= \delta D_{it} + u_i + \varepsilon_{it} \\ \bar{Y}_i &= \delta \bar{D}_i + u_i + \bar{\varepsilon}_i \\ (Y_{it} - \bar{Y}_i) &= (\delta D_{it} - \delta \bar{D}_i) + (u_i - u_i) + (\varepsilon_{it} - \bar{\varepsilon}_i) \\ \ddot{Y}_{it} &= \delta \ddot{D}_{it} + \ddot{\varepsilon}_{it} \end{aligned}$$

Where'd the unobserved heterogeneity go?! It was deleted when we time demeaned the data. And as we said, including individual fixed effects does this time demeaning automatically so that you don't have to go to the actual trouble of doing it yourself manually.¹²⁸

So how do we precisely do this form of estimation? There are three ways to implement the fixed effects (within) estimator. They are:

1. Demean and regress \ddot{Y}_{it} on \ddot{D}_{it} (need to correct degrees of freedom)
2. Regress Y_{it} on D_{it} and unit dummies (dummy variable regression)
3. Regress Y_{it} on D_{it} with canned fixed effects routine in Stata

```
. xtreg y d, fe i(PanelID)
```

More on the Stata implementation later at the end of this chapter. We'll review an example from my research and you'll estimate a POLS, a FE and a demeaned OLS model on real data so that you can see how to do this.

Identifying Assumptions We kind of reviewed the assumptions necessary to identify δ with our fixed effects (within) estimator when we walked through that original DAG, but let's supplement that DAG intuition with some formality. The main identification assumptions are:

1. $E[\varepsilon_{it}|D_{i1}, D_{i2}, \dots, D_{iT}, u_i] = 0; t = 1, 2, \dots, T$
 - This means that the regressors are strictly exogenous conditional on the unobserved effect. This allows D_{it} to be arbitrarily related to u_i , though. It only concerns the relationship between D_{it} and ε_{it} , not D_{it} 's relationship to u_i .
2. $\text{rank}\left(\sum_{t=1}^T E[\ddot{D}'_{it} \ddot{D}_{it}]\right) = K$
 - It shouldn't be a surprise to you by this point that we have a rank condition, because even when we were working with the simpler linear models, the estimated coefficient was always a

¹²⁸ Though feel free to do it if you want to convince yourself that they are numerically equivalent, probably just starting with a bivariate regression for simplicity.

scaled covariance, where the scaling was by a variance term. Thus regressors must vary over time for at least some i and not be collinear in order that $\hat{\delta} \approx \delta$.

The properties of the estimator under assumptions 1-2 are that $\hat{\delta}_{FE}$ is consistent ($\text{plim}_{N \rightarrow \infty} \hat{\delta}_{FE,N} = \delta$) and $\hat{\delta}_{FE}$ is unbiased conditional on \mathbf{D}

I only briefly mention inference. But the standard errors in this framework must be “clustered” by panel unit (e.g., individual) to allow for correlation in the ε_{it} 's for the same person i over time. In Stata, this is implemented as follows:

```
. xtreg y d , fe i(PanelID) cluster(PanelID)
```

This yields valid inference so long as the number of clusters is “large”.¹²⁹

Caveat #1: Fixed Effects Cannot Address Reverse Causality But, there are still things that fixed effects (within) estimators cannot solve. For instance, let's say we regressed crime rates onto police spending per capita. Becker [1968] argues that increases in the probability of arrest, usually proxied by police per capita or police spending per capita, will reduce crime. But at the same time, police spending per capita is itself a function of crime rates. This kind of reverse causality problem shows up in most panel models when regressing crime rates onto police. For instance, see Cornwell and Trumbull [1994], Table 3, column 2 (Figure 76). Focus on the coefficient on “POLICE”. The dependent variable is crime rates by county in North Carolina for a panel, and they find a *positive* correlation between police and crime rates. Does this mean the more police in an area *causes* higher crime rates? Or does it likely reflect the reverse causality problem?

Traditionally, economists have solved this kind of reverse causality problem by using instrumental variables. Examples include Evans and Owens [2007] and Draca et al. [2011]. I produce one example from Draca et al. [2011]. In this study, the authors used as an instrument in which police were deployed in response to terrorist attacks in London in a program called Operation Theseus. The authors present both a pooled OLS estimate (which is positive on the effect that police have on crime) and the 2SLS estimate (which is negative, consistent with Becker's hypothesis). See Figure 77.

So, one situation in which you wouldn't want to use panel fixed effects is if you have reverse causality or simultaneity bias. And specifically when that reverse causality is very strong in observational data. This would technically violate the DAG, though, that we presented at the start of the chapter. Notice that if we had reverse causality, then $Y \rightarrow D$, which is explicitly ruled out by this theoretical model

¹²⁹ In my experience, when an econometrician is asked how large is large, they say “the size of your data”. But that said, there is a small clusters literature and usually it's thought that fewer than 30 clusters is too small (as a rule of thumb). So it may be that having around 30-40 clusters is sufficient for the approaching of infinity. This will usually hold in most panel applications such as US states or individuals in the NSLY, etc.

NOTES

TABLE 3.—RESULTS FROM ESTIMATION
(standard errors in parentheses)

	Between	Within	2SLS (fixed effects)	2SLS (no fixed effects)
CONSTANT	-2.097 (2.822)			-3.719 (8.189)
P_A	-0.648 (0.088)	-0.355 (0.032)	-0.455 (0.618)	-0.507 (0.251)
P_C	-0.528 (0.067)	-0.282 (0.021)	-0.336 (0.371)	-0.530 (0.110)
P_P	0.297 (0.231)	-0.173 (0.032)	-0.196 (0.200)	0.200 (0.343)
S	-0.236 (0.174)	-0.00245 (0.02612)	-0.0298 (0.0300)	-0.218 (0.185)
<i>POLICE</i>	0.364 (0.060)	0.413 (0.027)	0.504 (0.617)	0.419 (0.218)
<i>DENSITY</i>	0.168 (0.077)	0.414 (0.283)	0.291 (0.785)	0.226 (0.103)
<i>PERCENT</i>	-0.0951	0.627	0.888	-0.145
<i>YOUNG MALE</i>	(0.1576)	(0.364)	(0.139)	(0.336)
<i>WCON</i>	0.195 (0.210)	-0.0378 (0.0391)	-0.0358 (0.0467)	0.329 (0.279)
<i>WTUC</i>	-0.196 (0.170)	0.0455 (0.0190)	0.0398 (0.0282)	-0.197 (0.197)
<i>WTRD</i>	0.129 (0.278)	-0.0205 (0.0405)	-0.0196 (0.0426)	0.0293 (0.3240)
<i>WFIR</i>	0.113 (0.220)	-0.00390 (0.02806)	-0.00700 (0.03270)	0.0506 (0.3224)
<i>WSER</i>	-0.106 (0.163)	0.00888 (0.01913)	0.00600 (0.02536)	-0.127 (0.176)
<i>WMFG</i>	-0.0249 (0.1339)	-0.360 (0.112)	-0.406 (0.217)	-0.0493 (0.1672)
<i>WFED</i>	0.156 (0.287)	-0.309 (0.176)	-0.273 (0.296)	0.170 (0.327)
<i>WSTA</i>	-0.284 (0.256)	0.0529 (0.114)	-0.0129 (0.2599)	-0.181 (0.300)
<i>WLOC</i>	0.0103 (0.4635)	0.182 (0.118)	0.136 (0.165)	0.0237 (0.5187)
<i>WEST</i>	-0.229 (0.108)			-0.198 (0.117)
<i>CENTRAL</i>	-0.164 (0.064)			-0.173 (0.067)
<i>URBAN</i>	-0.0346 (0.1324)			-0.0874 (0.1508)
<i>PERCENT</i>	0.148			0.174
<i>MINORITY</i>	(0.049)			(0.057)
s.e.	0.216	0.137	0.141	0.224

Figure 76: Table 3 from Cornwell and Trumbull [1994]

contained in the DAG. But obviously, in the police - crime example, that DAG would be inappropriate, and any amount of reflection on the problem should tell you that that DAG is inappropriate. Thus it requires, as I've said repeatedly, some careful reflection, and writing out exactly what the relationship is between the treatment variables and the outcome variables in a DAG can help you develop a credible identification strategy.

Caveat #2: Fixed Effects Cannot Address Time-variant Unobserved Heterogeneity The second situation in which panel fixed effects don't buy you anything is if the unobserved heterogeneity is time varying. In this situation, the demeaning has simply demeaned an unobserved time-variant variable, which is then moved into the composite error term, and which since time demeaned \ddot{u}_{it} correlated with \ddot{D}_{it} ,

	OLS Estimates		IV Estimates		
	Levels (1)	Differences (2)	Full (3)	Split (4)	+Trends (5)
<i>Panel C. Structural form</i>					
ln(police hours)	0.785*** (0.053)				
$\Delta \ln(\text{police hours})$		-0.031 (0.051)	-0.641** (0.301)	-0.318*** (0.093)	-0.183*** (0.066)
Controls	Yes	Yes	Yes	Yes	Yes
Trends	No	No	No	No	Yes
Number of boroughs	32	32	32	32	32
Observations	3,328	1,664	1,664	1,664	1,664

Figure 77: Table 2 from Draca et al. [2011]

\ddot{D}_{it} remains endogenous. Again, look carefully at the DAG - panel fixed effects is only appropriate if u_i is unchanging. Otherwise it's just another form of omitted variable bias. So, that said, don't just blindly use fixed effects and think that it solves your omitted variable bias problem – in the same way that you shouldn't use matching just because it's convenient to do. You need a DAG, based on an actual economic model, which will allow you to build the appropriate research design. Nothing substitutes for careful reasoning and economic theory, as they are the necessary conditions for good research design.

Example: Returns to Marriage and Unobserved Heterogeneity

When might this be true? Let's use an example from Cornwell and Rupert [1997] in which the authors attempt to estimate the causal effect of marriage on earnings. It's a well known stylized fact that married men earn more than unobserved men, even controlling for observables. But the question is whether that correlation is causal, or whether it reflects unobserved heterogeneity, or selection bias.

So let's say that we had panel data on individuals. These individuals i are observed for four periods t . We are interested in the following equation:¹³⁰

$$Y_{it} = \alpha + \delta M_{it} + \beta X_{it} + A_i + \gamma_i + \varepsilon_{it}$$

Let the outcome be their wage Y_{it} observed in each period, and which changes each period. Let wages be a function of marriage which changes over time M_{it} , other covariates the change over time X_{it} , race and gender which do not change over the panel period A_i , and an unobserved variable we call unobserved ability γ_i . This could be intelligence, non-cognitive ability, motivation, or some other unobserved confounder. The key here is that it is unit-specific, unobserved, and time-invariant. The ε_{it} is the unobserved determinants of wages which are assumed to be uncorrelated with marriage and other covariates.

¹³⁰ We use the same notation as used in their paper, as opposed to the \ddot{Y} notation presented earlier.

Cornwell and Rupert [1997] estimate both a feasible generalized least squares model and three fixed effects models (each of which includes different time-varying controls). The authors call the fixed effects regression a “within” estimator, because it uses the within unit variation for eliminating the confounding. Their estimates are presented in Figure 78.

TABLE II
Estimated Wage Regressions
(Standard Errors in Parentheses)

Variable	(1) FGLS	(2) Within	(3) Within	(4) Within
Married	0.083 (0.022)	0.056 (0.026)	0.051 (0.026)	0.033 (0.028)
Divorced	0.064 (0.033)	0.062 (0.036)	0.057 (0.036)	0.040 (0.038)
Years Married				-0.005 (0.006)
Years Married ²				-0.0003 (0.0003)
Years Divorced				-0.014 (0.008)
Experience	0.027 (0.004)	0.027 (0.004)	0.024 (0.004)	0.021 (0.005)
Experience ²	-0.001 (0.0001)	-0.001 (0.0002)	-0.001 (0.0002)	-0.001 (0.0002)
Tenure			0.013 (0.004)	0.011 (0.004)
Tenure ²			-0.0006 (0.0002)	-0.0005 (0.0002)
South	-0.091 (0.019)	-0.121 (0.034)	-0.117 (0.034)	-0.118 (0.034)
Urban	0.137 (0.017)	0.057 (0.024)	0.059 (0.024)	0.059 (0.024)
Union	0.109 (0.015)	0.106 (0.018)	0.102 (0.018)	0.103 (0.018)
Dependents	0.052 (0.017)	0.052 (0.019)	0.048 (0.019)	0.047 (0.020)
No High School	-0.325 (0.057)			
Some High School	-0.148 (0.032)			
Some College	0.091 (0.028)			
College Grad	0.278 (0.034)			
Post-College	0.322 (0.041)			
Standard error	0.215	0.212	0.212	0.211
χ^2_{27}	111.9			

Figure 78: Table 2 from Cornwell and Rupert [1997]

Notice that the FGLS (column 1) finds a strong marriage premium of around 8.3%. But, once we begin estimating fixed effects models, the effect gets smaller and less precise. The inclusion of marriage characteristics, such as years married and job tenure, causes the coefficient on marriage to fall by around 60% from the FGLS estimate, and is no longer statistically significant at the 5% level.

One of the interesting features of this analysis is the effect of dependents on wages. Even under the fixed effects estimation, the

relationship between dependents and wages is positive, robust and statistically significant. The authors explore this in more detail by including interactions of marriage variables with dependents (Figure 79). Here we see that the coefficient on marriage falls and is no longer statistically significant, but there still exists a positive effect of dependents on earnings.

TABLE III
Dependents and the Returns to Marriage
(Standard Errors in Parentheses)

Variable	(1) Within	(2) Within
Married	0.071 (0.026)	0.027 (0.038)
Divorced	0.008 (0.049)	-0.033 (0.058)
Years Married		0.011 (0.012)
Years Married ²		-0.001 (0.0007)
Years Divorced		-0.013 (0.011)
Experience	0.024 (0.004)	0.021 (0.005)
Experience ²	-0.001 (0.002)	-0.001 (0.0002)
Tenure	0.013 (0.004)	0.011 (0.004)
Tenure ²	-0.0006 (0.0002)	-0.0005 (0.0002)
South	-0.113 (0.034)	-0.109 (0.034)
Urban	0.061 (0.023)	0.061 (0.024)
Union	0.101 (0.018)	0.110 (0.018)
Dependents	0.292 (0.076)	0.281 (0.076)
Married × Dependents	-0.266 (0.077)	-0.232 (0.087)
Divorced × Dependents	-0.156 (0.095)	-0.124 (0.103)
Years Married × Dependents		-0.013 (0.012)
Years Married ² × Dependents		0.001 (0.0008)
Years Divorced × Dependents		-0.001 (0.012)
Standard error	0.213	0.211

Figure 79: Table 3 from Cornwell and Rupert [1997]

Stata example: Survey of Adult Service Providers

Next I'd like to introduce a Stata exercise based on data collection for my own research: a survey of sex workers. You may or may not know this, but the Internet has had a profound effect on sex markets.

It has moved women indoor from the streets while simultaneously breaking the link with pimps. It has increased safety and anonymity, too, which has had the effect of causing new entrants. The marginal sex worker has more education and better outside options than traditional US sex workers [Cunningham and Kendall, 2011, Cornwell and Cunningham, 2016]. The Internet, in sum, caused the marginal sex worker to shift towards women more sensitive to detection, harm and arrest.

In 2008 and 2009, I surveyed (with Todd Kendall) approximately 700 US Internet-mediated sex workers. The survey was a basic labor market survey; I asked them about their illicit and legal labor market experiences, and demographics. The survey had two parts: a “static” provider-specific section and a “panel” section. The panel section asked respondents to share information about each of the last 4 session with clients.¹³¹

I have created a shortened version of the dataset and uploaded it to my website. It includes a few time-invariant provider characteristics, such as race, age, marital status, years of schooling and body mass index, as well as several time-variant session-specific characteristics including the log of the hourly price, the log of the session length (in hours), characteristics of the client himself, whether a condom was used in any capacity during the session, whether the client was a “regular”, etc.

In this exercise, you will estimate three types of models: a pooled OLS model, a fixed effects (FE) and a demeaned OLS model. The model will be of the following form:

$$\begin{aligned} Y_{is} &= \beta_i X_i + \gamma_{is} Z_{is} + u_i + \varepsilon_{is} \\ \ddot{Y}_{is} &= \gamma_{is} \ddot{Z}_{is} + \ddot{\eta}_{is} \end{aligned}$$

where u_i is both unobserved and correlated with Z_{is} .

The first regression model will be estimated with pooled OLS and the second model will be estimated using both fixed effects and OLS. In other words, I’m going to have you estimate the model using the `xtreg` function with individual fixed effects, as well as demean the data manually and estimate the demeaned regression using `reg`.

Notice that the second regression has a different notation on the dependent and independent variable; it represents the fact that the variables are columns of *demeaned* variables. Thus $\ddot{Y}_{is} = Y_{is} - \bar{Y}_i$. Secondly, notice that the time-invariant X_i variables are missing from the second equation. Do you understand why that is the case? These variables have also been demeaned, but since the demeaning is across time, and since these time-invariant variables do not change over time, the demeaning deletes them from the expression. Notice, also, that the unobserved individual specific heterogeneity, u_i , has

¹³¹ Technically, I asked them to share about the last five sessions, but for this exercise, I have dropped the fifth due to low response rates on the fifth session.

disappeared. It has disappeared for the same reason that the X_i terms are gone – because the mean of u_i over time is itself, and thus the demeaning deletes it.

To estimate these models, type the following lines into Stata:

```
. scuse sasp_panel, clear
. tset id session
. foreach x of varlist lnw age asq bmi hispanic black other asian schooling cohab married divorced //
separated age_cl unsafe llength reg asq_cl appearance_cl provider_second asian_cl black_cl hispanic_cl //
othrace_cl hot massage_cl {
drop if 'x'==.
}
. bysort id: gen s=_N
. keep if s==4
. foreach x of varlist lnw age asq bmi hispanic black other asian schooling cohab married divorced //
separated age_cl unsafe llength reg asq_cl appearance_cl provider_second asian_cl black_cl hispanic_cl //
othrace_cl hot massage_cl {
egen mean_`x'=mean(`x'), by(id)
gen demean_`x'=`x' - mean_`x'
drop mean*
}
. xi: reg lnw age asq bmi hispanic black other asian schooling cohab married divorced separated //
age_cl unsafe llength reg asq_cl appearance_cl provider_second asian_cl black_cl hispanic_cl //
othrace_cl hot massage_cl, robust

. xi: xtreg lnw age asq bmi hispanic black other asian schooling cohab married divorced separated //
age_cl unsafe llength reg asq_cl appearance_cl provider_second asian_cl black_cl hispanic_cl //
othrace_cl hot massage_cl, fe i(id) robust
. reg demean_lnw demean_age demean_asq demean_bmi demean_hispanic demean_black demean_other // 
demean_asian demean_schooling demean_cohab demean_married demean_divorced demean_separated // 
demean_age_cl demean_unsafe demean_llength demean_reg demean_asq_cl demean_appearance_cl // 
demean_provider_second demean_asian_cl demean_black_cl demean_hispanic_cl demean_othrace_cl // 
demean_hot demean_massage_cl, robust cluster(id)
```

Notice the first five commands created a balanced panel. Some of the respondents would leave certain questions blank, probably due to

concerns about anonymity and privacy. So we have dropped anyone who had missing values for the sake of this exercise. This leaves us with a balanced panel. You can see this yourself if after running those five lines you type `xtdescribe`.

I have organized the output into Table 31. There's a lot of interesting information in these three columns, some of which may surprise you if only for the novelty of the regressions. So let's talk about the statistically significant ones. The pooled OLS regressions, recall, do not control for unobserved heterogeneity, because by definition those are unobservable. So these are potentially biased by the unobserved heterogeneity, which is a kind of selection bias, but we will discuss them anyhow.

First, a simple scan of the second and third column will show that the fixed effects regression which included (not shown) dummies for the individual herself is equivalent to a regression on the demeaned data. This should help persuade you that the fixed effects and the demeaned (within) estimators are yielding the same coefficients.

But second, let's dig into the results. One of the first things we observe is that in the pooled POLS model, there is not a compensating wage differential detectable on having unprotected sex with a client.¹³² But, notice that in the fixed effects model, unprotected sex has a premium. This is consistent with Rosen [1986] who posited the existence of risk premia, as well as Gertler et al. [2005] who found risk premia for sex workers using panel data. Gertler et al. [2005], though, find a much larger premia of over 20% for unprotected sex, whereas I am finding only a mere 5%. This could be because a large number of the unprotected instances are fellatio, which carry a much lower risk of infection than unprotected receptive intercourse. Nevertheless, it is interesting that unprotected sex, under the assumption of strict exogeneity, appears to cause wages to rise by approximately 5%, which is statistically significant at the 10% level. Given an hourly wage of \$262, this amounts to a mere \$13 additional dollars per hour. The lack of a finding in the pooled OLS model seems to suggest that the unobserved heterogeneity was masking the effect.

Next we look at the session length. Note that I have already adjusted the price the client paid for the length of the session so that the outcome is a log wage, as opposed to a log price. As this is a log-log regression, we can interpret the coefficient on log length as an elasticity. When we use fixed effects, the elasticity increases from -0.308 to -0.435. The significance of this result, in economic terms, though, is that there appears to be "volume discounts" in sex work. That is, longer sessions are more expensive, but at a decreasing rate. Another interesting result is whether the client was a "regular" which meant that she had seen him before in another session. In our pooled OLS

¹³² There were three kinds of sexual encounter - vaginal receptive sex, anal receptive sex, and fellatio. Unprotected sex is coded as any sex act without a condom.

Depvar:	POLS	FE	Demeaned OLS
Unprotected sex with client of any kind	0.013 (0.028)	0.051* (0.028)	0.051* (0.026)
Ln(Length)	-0.308*** (0.028)	-0.435*** (0.024)	-0.435*** (0.019)
Client was a Regular	-0.047* (0.028)	-0.037** (0.019)	-0.037** (0.017)
Age of Client	-0.001 (0.009)	0.002 (0.007)	0.002 (0.006)
Age of Client Squared	0.000 (0.000)	-0.000 (0.000)	-0.000 (0.000)
Client Attractiveness (Scale of 1 to 10)	0.020*** (0.007)	0.006 (0.006)	0.006 (0.005)
Second Provider Involved	0.055 (0.067)	0.113* (0.060)	0.113* (0.048)
Asian Client	-0.014 (0.049)	-0.010 (0.034)	-0.010 (0.030)
Black Client	0.092 (0.073)	0.027 (0.042)	0.027 (0.037)
Hispanic Client	0.052 (0.080)	-0.062 (0.052)	-0.062 (0.045)
Other Ethnicity Client	0.156** (0.068)	0.142*** (0.049)	0.142*** (0.045)
Met Client in Hotel	0.133*** (0.029)	0.052* (0.027)	0.052* (0.024)
Gave Client a Massage	-0.134*** (0.029)	-0.001 (0.028)	-0.001 (0.024)
Age of provider	0.003 (0.012)	0.000 (.)	0.000 (.)
Age of provider squared	-0.000 (0.000)	0.000 (.)	0.000 (.)
Body Mass Index	-0.022*** (0.002)	0.000 (.)	0.000 (.)
Hispanic	-0.226*** (0.082)	0.000 (.)	0.000 (.)
Black	0.028 (0.064)	0.000 (.)	0.000 (.)
Other	-0.112 (0.077)	0.000 (.)	0.000 (.)
Asian	0.086 (0.158)	0.000 (.)	0.000 (.)
Imputed Years of Schooling	0.020** (0.010)	0.000 (.)	0.000 (.)
Cohabitating (living with a partner) but unmarried	-0.054 (0.036)	0.000 (.)	0.000 (.)
Currently married and living with your spouse	0.005 (0.043)	0.000 (.)	0.000 (.)
Divorced and not remarried	-0.021 (0.038)	0.000 (.)	0.000 (.)
Married but not currently living with your spouse	-0.056 (0.059)	0.000 (.)	0.000 (.)
N	1,028	1,028	1,028
Mean of dependent variable	5.57	5.57	0.00

Heteroskedastic robust standard errors in parenthesis clustered at the provider level. * p<0.10, ** p<0.05, *** p<0.01

Table 31: POLS, FE and Demeaned OLS Estimates of the Determinants of Log Hourly Price for a Panel of Sex Workers

model, regulars paid 4.7% less, but this shrinks slightly in our fixed effects model to 3.7% reductions. Economically, this could be lower because new clients pose risks that repeat customers do not pose. Thus, if we expect prices to move closer to marginal cost, the disappearance of some of the risk from the repeated session should lower price, which it appears to do.

Another factor related to price is the attractiveness of the client. Interestingly, this does not go in the direction we may have expected. One might expect that the more attractive the client, the *less* he pays. But in fact it is the opposite. Given other research that finds beautiful people earn more money [Hamermesh and Biddle, 1994], it's possible that sex workers are price discriminating. That is, when they see a handsome client, they deduce he earns more, and therefore charges him more. This result does not hold up when including fixed effects, though, suggesting that it is due to unobserved heterogeneity, at least in part.

Similar to unprotected sex, a second provider present has a positive effect on price which is only detectable in the fixed effects model. Controlling for unobserved heterogeneity, the presence of a second provider increases prices by 11.3%. We also see that she discriminates against clients of "other" ethnicity who pay 14.2% more than White clients. There's a premium associated with meeting in a hotel which is considerably smaller when controlling for provider fixed effects by almost a third. This positive effect, even in the fixed effects model, may simply represent the higher costs associated with meeting in a hotel room. The other coefficients are not statistically significant.

Many of the time-invariant results are also interesting, though. For instance, perhaps not surprisingly, women with higher BMI earn less. Hispanics earn less than White sex workers. And women with more schooling earn more, something which is explored in greater detail in Cunningham and Kendall [2016].

Conclusion In conclusion, we have been exploring the usefulness of panel data for estimating causal effects. We noted that the fixed effects (within) estimator is a very useful method for addressing a very specific form of endogeneity, with some caveats. First, it will eliminate any and all unobserved and observed time-invariant covariates correlated with the treatment variable. So long as the treatment and the outcome varies over time, and strict exogeneity, then the fixed effects (within) estimator will identify the causal effect of the treatment on some outcome.

But this came with certain qualifications. For one, the method couldn't handle *time variant* unobserved heterogeneity. It's thus the burden of the researcher to determine which type of unobserved

heterogeneity problem they face, but if they face the latter, then the panel methods reviewed here are not unbiased and consistent. Second, when there exists strong reverse causality pathways, then panel methods are biased. Thus, we cannot solve the problem of simultaneity, such as what Wright faced when estimating the price elasticity of demand, using the fixed effects (within) estimator. Most likely, we are going to have to move into a different framework when facing that kind of problem.

Still, many problems in the social sciences may credibly be caused by a time-invariant unobserved heterogeneity problem, in which case the fixed effects (within) panel estimator is useful and appropriate.

Differences-in-differences

Introduction

In 2002, Craigslist opened a new section on its front page called “erotic services” in San Francisco, California. The section would end up being used by sex workers exclusively to advertise to and solicit clients. Sex workers claimed it made them safer, because instead of working on street corners and for pimps, they could solicit indoors from their computers, which as a bonus, also gave them the chance to learn more about the men contacting them. But activists and law enforcement worried that it was facilitating sex trafficking and increasing violence against women. Which was it? Was erotic services (ERS) making women safer, or was it placing them in harm’s way?

This is ultimately an empirical question. We want to know the effect of ERS on female safety, but the fundamental problem of causal inference says that we can’t know what effect it had because we are missing the data necessary to make the calculation. That is,

$$E[\delta] = E[M^1 - M^0]$$

where M^1 is women murdered in a world where San Francisco has ERS, and M^0 is women murdered in a world where San Francisco does not have ERS *at the exact same moment in time*. In 2002, only the first occurred, as the second was a counterfactual. So how do we proceed?

The standard way to evaluate interventions such as this is the standard differences-in-differences strategy, or DD.¹³³ DD is basically a version of panel fixed effects, but can also be used with repeated cross-sections. Let’s look at this example using some tables, which hopefully will help give you an idea of the intuition behind DD, as well as some of its identifying assumptions.

Let’s say that the intervention is erotic services, or E , and we want to know the causal effect of E on female murders M . Couldn’t we just compare San Francisco murders in, say, 2003 with some other city, like Waco, Texas, where the author lives? Let’s look at that.

¹³³ You’ll sometimes see the acronyms DiD, Diff-in-diff, or even DnD.

Cities	Outcome
San Francisco	$M = SF + E$
Waco, Texas	$M = W$

Table 32: Compared to what? Different cities

where SF is an unobserved San Francisco fixed effect and W is a Waco fixed effect. When we make a simple comparison between Waco and San Francisco, we get a causal effect equalling $E + SF - W$. Thus the simple difference is biased because of W and SF . Notice that the $SF - W$ term is akin to our selection bias term in the decomposition of the simpler difference in outcomes. It's the underlying differences in murder rates between the two cities in a world where neither gets treated. So if our goal is to get an unbiased estimate of E , then that simple difference won't work unless W and SF are the same.

But what if we compared San Francisco to itself? Say compared it in 2003 to two years earlier in 2001? Let's look at that simple before and after difference. Again, this doesn't lead to an unbiased

Cities	Time	Outcome
San Francisco	Before	$M = SF$
	After	$M = SF + T + E$

Table 33: Compared to what? Before and After

estimate of E , even if it does eliminate the fixed effect. That's because such differences can't control for or net out natural changes in the murder rate over time. I can't compare San Francisco before and after ($T + E$) because of T which is a kind of omitted variable bias. If we could control for T , then it'd be fine, though.

The intuition of the DD strategy is simple: all you do is combine these two simpler approaches so that you can eliminate both the selection bias and the effect of time. Let's look at it in the following table. The first difference, D_1 , does the simple before and after

Cities	Time	Outcome	D_1	D_2
San Francisco	Before	$M = SF$		
	After	$M = SF + T + E$	$T + E$	E
Waco	Before	$M = W$		
	After	$M = W + T$	T	

Table 34: Compared to what? Subtract each city's differences

difference. This ultimately eliminates the unit specific fixed effects. Then, once those differences are made, we difference the differences (hence the name) to get the unbiased estimate of E . But there's a couple of key assumptions with a standard DD model. First, we are

assuming that there is no time-variant city specific unobservables. Nothing unobserved in San Francisco that is changing over time that *also* determines murders. And secondly, we are assuming that T is the same for all units. This second assumption is called the parallel trends assumption, which I'll discuss in more detail later.

DD is a powerful, yet amazingly simple, strategy. It is a kind of panel estimator in the sense that it utilizes repeated observations on the same unit to eliminate the unobserved heterogeneity confounding the estimate of the treatment effect. But here we treat it separately because of the amount of focus it has gotten separately in the literature.

Background

You see traces of this kind of strategy in Snow's cholera study, though technically Snow only did a simple difference. He just had every reason to believe that absent the treatment, the two parts of London would've had similar underlying cholera rates since the two groups were so similar ex ante. The first time I ever saw DD in its current form was [Card and Krueger \[1994\]](#), a famous minimum wage study. This was a famous study primarily because of its use of an explicit counterfactual for estimation. Suppose you are interested in the effect of minimum wages on employment. Theoretically, you might expect that in competitive labor markets, an increase in the minimum wage would move us up a downward sloping demand curve causing employment to fall. But [Card and Krueger \[1994\]](#) was interested in quantifying this, and approached it furthermore as though it was purely an empirical question.

Their strategy was to do a simple DD between two neighboring states - a strategy we would see again in minimum wage research with [Dube et al. \[2010\]](#). New Jersey was set to experience an increase in the state minimum wage from \$4.25 to \$5.05, but neighboring Pennsylvania's minimum wage was staying at \$4.25 (see Figure 8o).

They surveyed about 400 fast food stores both in New Jersey and Pennsylvania before and after the minimum wage increase. This was used to measure the outcomes they cared about (i.e., employment).

Let Y_{ist}^1 be employment at restaurant i , in state s , at time t with a high minimum wage, and let Y_{ist}^0 be employment at restaurant i , state s , time t with a low minimum wage. As we've said repeatedly through this book, we only see one or the other because the switching equation selects one or the other based on the treatment assignment. But, we can assume then that

$$E[Y_{ist}^0 | s, t] = \gamma_s + \tau_t$$

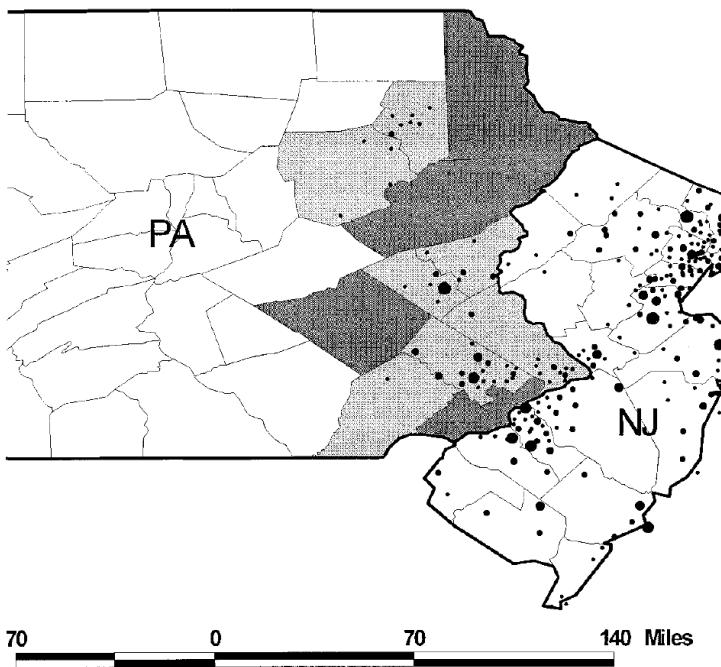


Figure 8o: NJ and PA

. In the absence of a minimum wage change, in other words, employment in a state will be determined by the sum of a time-invariant state fixed effect, γ_s , that is idiosyncratic to the state, and a time effect τ_t that is common across all states.

Let D_{st} be a dummy for high-minimum wage states and periods. Under the conditional independence assumption, we can write out the average treatment effect as

$$E[Y_{ist}^1 - Y_{ist}^0 | s, t] = \delta$$

and observed employment can be written as

$$Y_{ist} = \gamma_s + \tau_t + \delta D_{st} + \varepsilon_{ist}$$

Figure 81 shows the distribution of wages in November 1992 after the minimum wage hike. As can be seen, the minimum wage hike was binding evidenced by the mass of wages at the minimum wage in New Jersey.

Now how do we take all this information and precisely calculate the treatment effect? One way is to do what we did earlier in our San Francisco and Waco example: compute before and after differences for each state, and then difference those differences.

In New Jersey:

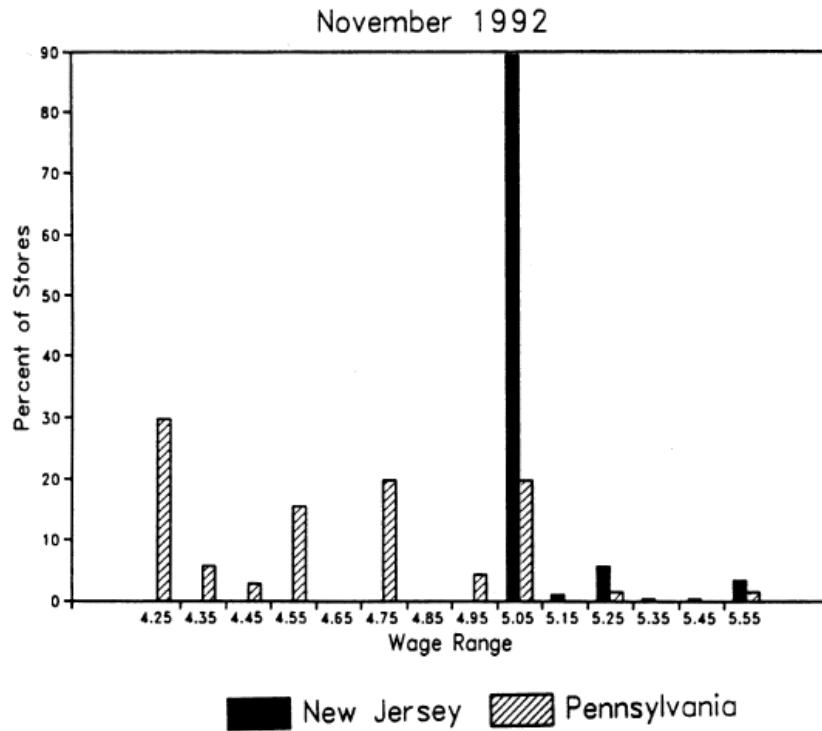


Figure 81: Distribution of wages for NJ and PA in November 1992

- Employment in February is

$$E(Y_{ist}|s = NJ, t = Feb) = \gamma_{NJ} + \lambda_{Feb}$$

- Employment in November is:

$$E(Y_{ist}|s = NJ, t = Nov) = \gamma_{NJ} + \lambda_{Nov} + \delta$$

- Difference between November and February

$$E(Y_{ist}|s = NJ, t = Nov) - E(Y_{ist}|s = NJ, t = Feb) = \lambda_N - \lambda_F + \delta$$

And in Pennsylvania:

- Employment in February is

$$E(Y_{ist}|s = PA, t = Feb) = \gamma_{PA} + \lambda_{Feb}$$

- Employment in November is:

$$E(Y_{ist}|s = PA, t = Nov) = \gamma_{PA} + \lambda_{Nov}$$

- Difference between November and February

$$E(Y_{ist}|s = PA, t = Nov) - E(Y_{ist}|s = PA, t = Feb) = \lambda_N - \lambda_F$$

Once we have those two before and after differences, we simply difference them each to net out the time effects. The DD strategy amounts to comparing the change in employment in NJ to the change in employment in PA. The population DD are:

$$\begin{aligned}\hat{\delta} &= \left(E(Y_{ist}|s = NJ, t = Nov) - E(Y_{ist}|s = NJ, t = Feb) \right) \\ &\quad - \left(E(Y_{ist}|s = PA, t = Nov) - E(Y_{ist}|s = PA, t = Feb) \right) \\ &= (\lambda_N - \lambda_F + \delta) - (\lambda_N - \lambda_F) \\ &= \delta\end{aligned}$$

This is estimated using the sample analog of the population means (see Figure 82).

Variable	Stores by state		
	PA (i)	NJ (ii)	Difference, NJ – PA (iii)
1. FTE employment before, all available observations	23.33 (1.35)	20.44 (0.51)	-2.89 (1.44)
2. FTE employment after, all available observations	21.17 (0.94)	21.03 (0.52)	-0.14 (1.07)
3. Change in mean FTE employment	-2.16 (1.25)	0.59 (0.54)	2.76 (1.36)

Figure 82: Simple DD using sample averages

What made this study so controversial was less its method and more its failure to find the negative effect on employment predicted by a neoclassical perfect competition model. In fact, not only did employment *not* fall; their DD showed it rose relative to the counterfactual. This paper started a new wave of studies on the minimum wage, which continues to this day.¹³⁴

Simple differencing is one way to do it, but it's not the only way to do it. We can also directly estimate this using a regression framework. The advantages of that is that we can control for other variables which may reduce the residual variance (leading to smaller standard errors), it's easy to include multiple time periods, and we can study treatments with different treatment intensity (e.g., varying increases in the minimum wage for different states). The typical regression model we estimate is:

$$Y_{it} = \alpha + \beta_1 D_i + \beta_2 Post_t + \delta(D \times Post)_{it} + \tau_t + \sigma_s + \varepsilon_{st}$$

¹³⁴ A review of that literature is beyond the scope of this chapter, but you can find a relatively recent review by Neumark et al. [2014].

where D is a dummy whether the unit is in the treatment group or not, $Post$ is a post-treatment dummy, and the interaction is the DD coefficient of interest.

One way to build this is to have as a separate variable the date in which a unit (e.g., state) received the treatment, and then generate a new variable equalling the difference between the current date and the date of treatment. So for instance, say that the current date is 2001 and the treatment occurred in 2004. Then 2001-2004 equals -3. This new variable would be a re-centering of the time period such that each unit was given a date from the point it received the treatment. Then one could define the post-treatment period as all periods where the recentered variable exceeded zero for those treatment units.

In the Card and Krueger case, the equivalent regression would be:

$$Y_{its} = \alpha + \gamma NJ_s + \lambda d_t + \delta(NJ \times d)_{st} + \varepsilon_{its}$$

NJ is a dummy equal to 1 if the observation is from NJ, and d is a dummy equal to 1 if the observation is from November (the post period). This equation takes the following values

- PA Pre: α
- PA Post: $\alpha + \lambda$
- NJ Pre: $\alpha + \gamma$
- NJ Post: $\alpha + \gamma + \lambda + \delta$

The DD estimate: $(NJ \text{ Post} - NJ \text{ Pre}) - (PA \text{ Post} - PA \text{ Pre}) = \delta$. We can see this visually in Figure 83.

Notice that the regression identifies a vertical bar in the post-treatment period marked by the δ . What's important to notice is that algebraically this is only the actual treatment effect if the declining line for NJ is exactly equal to the declining line for PA. In other words, it's because of these parallel trends that the object identified by the regression equals in expectation the true parameter.

This gets to our key identifying assumption in DD strategies – the parallel trends assumption. This is simply an untestable assumption because as we can see, we don't know what would've happened to employment in New Jersey had they not passed the minimum wage because that is a counterfactual state of the world. Maybe it would've evolved the same as Pennsylvania, but maybe it wouldn't have too. We have no way of knowing.

Empiricists faced with this untestable assumption have chosen, therefore, to use deduction as a second best for checking the assumption. By which I mean, empiricists will reason that if the pre-treatment trends were parallel between the two groups, then

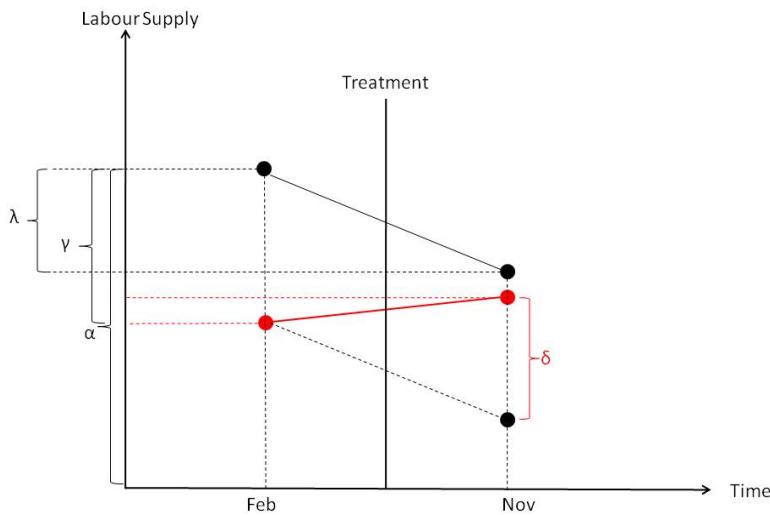


Figure 83: DD regression diagram

wouldn't it stand to reason that the post-treatment trends *would have too*? Notice, this is not a test of the assumption; rather, this is a test of a possible corollary of the assumption: checking the pre-treatment trends. I emphasize this because I want you to understand that checking the parallelism of the pre-treatment trends is *not* equivalent to proving that the post-treatment trends would've evolved the same. But given we see that the pre-treatment trends evolved similarly, it does give some confidence that the post-treatment would've too (absent some unobserved group specific time shock). That would look like this (see Figure 84): Including leads into the DD model is an easy way to check for the pre-treatment trends. Lags can be included to analyze whether the treatment effect changes over time after treatment assignment, too. If you did this, then the estimating regression equation would be:

$$Y_{its} = \gamma_s + \lambda_t + \sum_{\tau=-q}^{-1} \gamma_\tau D_{s\tau} + \sum_{\tau=0}^m \delta_\tau D_{s\tau} + x_{ist} + \varepsilon_{ist}$$

Treatment occurs in year 0. You include q leads or anticipatory effects and m leads or post treatment effects. Boom goes the dynamite.

Autor [2003] included both leads and lags in his DD model when he studied the effect of increased employment protection on the firms' use of temporary help workers. In the US, employers can usually hire and fire at will, but some state courts have made exceptions to this "employment at will" rule and have thus increased employment protection. The standard thing in this kind of analysis is to do what I said earlier and re-center the adoption year to 0. Autor [2003] then analyzed the effects of these exemptions on the use of tempo-

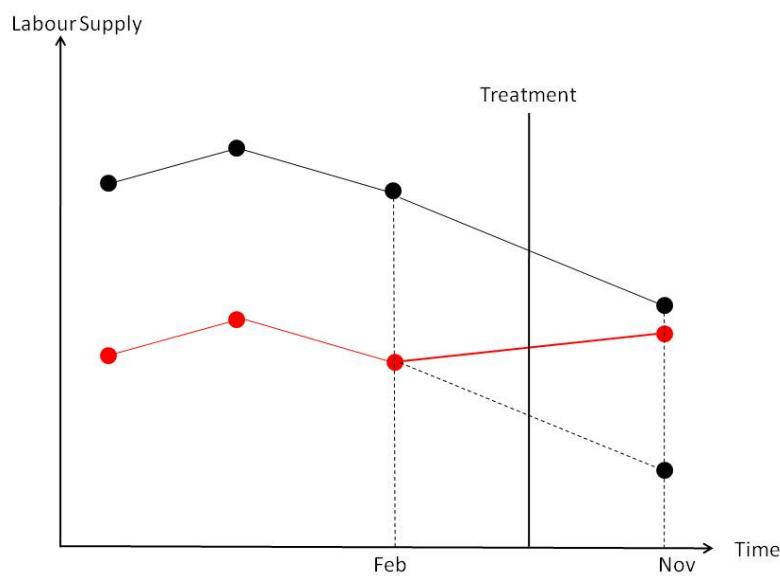


Figure 84: Checking the pre-treatment trends for parallelism

rary health workers. These results are shown in Figure 85. Notice

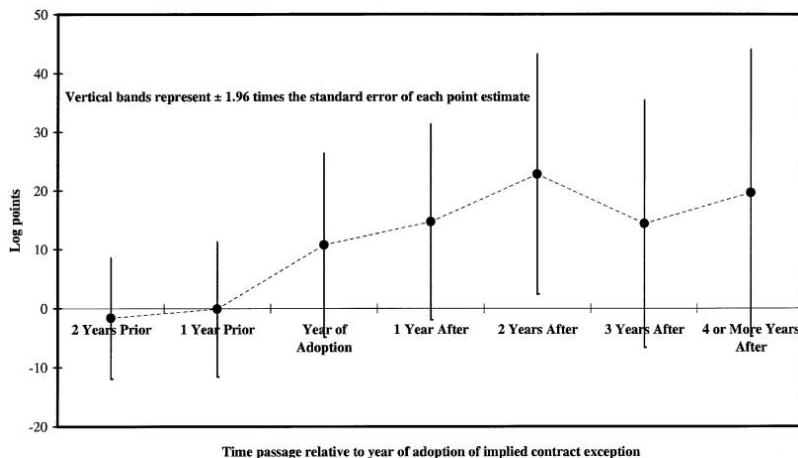


Figure 85: Autor [2003] leads and lags in dynamic DD model

that the leads are very close to 0. Thus, there is no evidence for anticipatory effects (good news for the parallel trends assumption). The lags show that the treatment effect is dynamic: it increases during the first few years, and then plateaus.

Inference Many papers using DD strategies use data from many years – not just 1 pre and 1 post treatment period like Card and

Krueger [1994]. The variables of interest in many of these setups only vary at a group level, such as the state, and outcome variables are often serially correlated. In Card and Krueger [1994], it is very likely for instance that employment in each state is not only correlated within the state but also serially correlated. Bertrand et al. [2004] point out that the conventional standard errors often severely underestimate the standard deviation of the estimators, and so standard errors are biased downward (i.e., incorrectly small). Bertrand et al. [2004] propose therefore the following solutions.

1. Block bootstrapping standard errors (if you analyze states the block should be the states and you would sample whole states with replacement for bootstrapping)
2. Clustering standard errors at the group level (in Stata one would simply add `, cluster(state)` to the regression equation if one analyzes state level variation)
3. Aggregating the data into one pre and one post period. Literally works if there is only one treatment data. With staggered treatment dates one should adopt the following procedure:
 - Regress Y_{st} onto state FE, year FE and relevant covariates
 - Obtain residuals from the treatment states only and divide them into 2 groups: pre and post treatment
 - Then regress the two groups of residuals onto a post dummy

Correct treatment of standard errors sometimes makes the number of groups very small: in Card and Krueger [1994], the number of groups is only 2. More common than not, researchers will use the second option (clustering the standard errors by group), though sometimes you'll see people do all three for robustness.

Threats to validity There are four threats to validity in a DD strategy. They are: (1) non-parallel trends; (2) compositional differences; (3) long-term effects vs. reliability; (4) functional form dependence. We discuss those now in order.

Regarding the violation of parallel trends, one way in which that happens is through endogenous treatments. Often policymakers will select the treatment and controls based on pre-existing differences in outcomes – practically guaranteeing the parallel trends assumption will be violated. One example of this is the “Ashenfelter dip”, named after Orley Ashenfelter, labor economist at Princeton. Participants in job trainings program often experience a “dip” in earnings just prior to entering the program. Since wages have a natural tendency to mean reversion, comparing wages of participants

and non-participants using DD leads to an upward biased estimate of the program effect. Another example is regional targeting, like when NGOs target villages that appear most promising, or worse off. This is a form of selection bias and violates parallel trends.

What can you do if you think the parallel trends assumption is violated? There's a variety of robustness checks that have become very common. They all come down to various forms of placebo analysis. For instance, you can look at the leads like we said. Or you can use a falsification test using data for an alternative control group, which I'll discuss in a moment. Or you can use a falsification test using alternative outcomes that shouldn't be affected by the treatment. For instance, if Craigslist's erotic services only helps female sex workers, then we might check that by estimating the same model against manslaughters and male murders – neither of which are predicted to be affected by ERS, but which would be affected by secular violence trends.

DDD The use of the alternative control group, though, is usually called the differences-in-differences-in-differences model, or DDD.¹³⁵ This was first introduced by Gruber [1994] in his study of maternity benefits. Before we dig into this paper, let's go back to our original DD table from the start of the chapter. What if we introduced city-specific time-variant heterogeneity? Then DD is biased. Let's see.

¹³⁵ Also called triple difference, DnDnD, or DiDiD.

Cities	Category	Period	Outcomes	D ₁	D ₂	D ₃ differences
San Francisco	Female murders	After	$SF + T + SF_t + f_t + \delta$	$T + SF_t + f_t + \delta$		
		Before	SF			
	Male murders	After	$SF + T + SF_t + m_t$		$\delta + f_t - m_t$	
		Before	SF	$T + SF_t + m_t$		
Waco	Female murders	After	$W + T + W_t + f_t$			δ
		Before	W	$T + W_t + f_t$		
	Male murders	After	$W + T + W_t + m_t$		$f_t - m_t$	
		Before	W	$T + W_t + m_t$		

The way that you read this table is as follows. Female murders in San Francisco are determined by some San Francisco fixed effect in the before period, and that same San Francisco fixed effect in the after period plus a time trend T , a San Francisco specific time trend, a trend in female murders separate from the national trend shaping all crimes, and the erotic services platform δ . When we difference this we get

$$T + SF_t + f_t + \delta$$

Now in the normal DD, we would do the same before and after differencing for Waco female murders, which would be

$$T + W_t + f_t$$

And if we differenced these two, we'd get

$$SF_t - W_t + \delta$$

This is the familiar selection bias term – the DD estimator would isolate the treatment effect plus the selection bias, and thus we couldn't know the effect itself.

The logic of the DDD strategy is to use a within-city comparison group that experiences the same city-specific trends, as well as its own crime-specific trend, and use these within-city controls to net them out. Go through each difference to confirm that at the third difference, you have isolated the treatment effect, δ . Note that while this seems to have solved the problem, it came at a cost, which is more parallel trends assumptions. That is, now we require that female murders have a common trend, the entire country have a common trend, and each city have a common trend. We also require that these crime outcomes be additive, otherwise the differencing would not eliminate the components from the analysis.

[Gruber \[1994\]](#) does this exact kind of triple differencing in his original maternity mandate paper. Here are his main results in [Figure 86](#):

These kinds of simple triple differencing are useful because they explain the intuition behind triple differencing, but in practice you will usually run regressions of the following form:

$$\begin{aligned} Y_{ijt} = & \alpha + \beta_1 X_{ijt} + \beta_2 \tau_t + \beta_3 \delta_j + \beta_4 D_i + \beta_5 (\delta \times \tau)_{jt} \\ & + \beta_6 (\tau \times D)_{ti} + \beta_7 (\delta \times D)_{ij} + \beta_8 (\delta \times \tau \times D)_{ijt} + \varepsilon_{ijt} \end{aligned}$$

where in this representation, the parameter of interest is β_8 . There's a few things I want to bring to your attention. First, notice the additional subscript, j . This j indexes whether it's the main category of interest (e.g., female murders) or the within-city comparison group (e.g., male murders).

But sometimes, it is sufficient just to use your DD model and use it to examine the effect of the treatment on a placebo as a falsification. For instance, [Cheng and Hoekstra \[2013\]](#) examined the effect of castle doctrine gun laws on homicides as their main results, but they performed placebo analysis by also looking at the law's effect on grand theft auto. [Auld and Grootendorst \[2004\]](#) estimated standard "rational addiction" models from [Becker and Murphy \[1988\]](#) on outcomes that could not possibly be considered addictive, such as

TABLE 3—DDD ESTIMATES OF THE IMPACT OF STATE MANDATES
ON HOURLY WAGES

Location/year	Before law change	After law change	Time difference for location
<i>A. Treatment Individuals: Married Women, 20–40 Years Old:</i>			
Experimental states	1.547 (0.012) [1,400]	1.513 (0.012) [1,496]	−0.034 (0.017)
Nonexperimental states	1.369 (0.010) [1,480]	1.397 (0.010) [1,640]	0.028 (0.014)
Location difference at a point in time:	0.178 (0.016)	0.116 (0.015)	
Difference-in-difference:		−0.062 (0.022)	
<i>B. Control Group: Over 40 and Single Males 20–40:</i>			
Experimental states	1.759 (0.007) [5,624]	1.748 (0.007) [5,407]	−0.011 (0.010)
Nonexperimental states	1.630 (0.007) [4,959]	1.627 (0.007) [4,928]	−0.003 (0.010)
Location difference at a point in time:	0.129 (0.010)	0.121 (0.010)	
Difference-in-difference:		−0.008: (0.014)	
DDD:		−0.054 (0.026)	

Figure 86: Gruber [1994] Table 3

eggs and milk. Since they find evidence for addiction with these models, they argued that the identification strategy that authors had been using previously to evaluate the rational addiction model were flawed. And then there is the networks literature. Several studies found significant network effects on outcomes like obesity, smoking, alcohol use and happiness, leading many researchers to conclude that these kinds of risk behaviors were “contagious” through peer effects. Cohen-Cole and Fletcher [2008] used similar models and data to study network effects for things that *couldn’t* be transmitted between peers – acne, height, and headaches – in order to show that the models’ research designs were flawed.

DD can be applied to repeated cross-sections, as well as panel data. But one of the risks of working with the repeated cross-section is that unlike panel data (e.g., individual level panel data), repeated cross-sections run the risk of compositional changes. Hong [2013] used repeated cross-sectional data from the Consumer Expenditure Survey (CEX) containing music expenditure and internet use for a random sample of households. The authors’ study exploited the emergence of Napster, the first file sharing software widely used by Internet users, in June 1999 as a natural experiment. The study compared Internet users and Internet non-users before and after emergence of Napster. Figure 87 shows the main results. Notice that as the Internet diffusion increased, music expenditure for the Internet user group declined – as did for the non-user group – suggesting that Napster was causing people to substitute away from music purchases towards file sharing.

Figure 1: Internet Diffusion and Average Quarterly Music Expenditure in the CEX

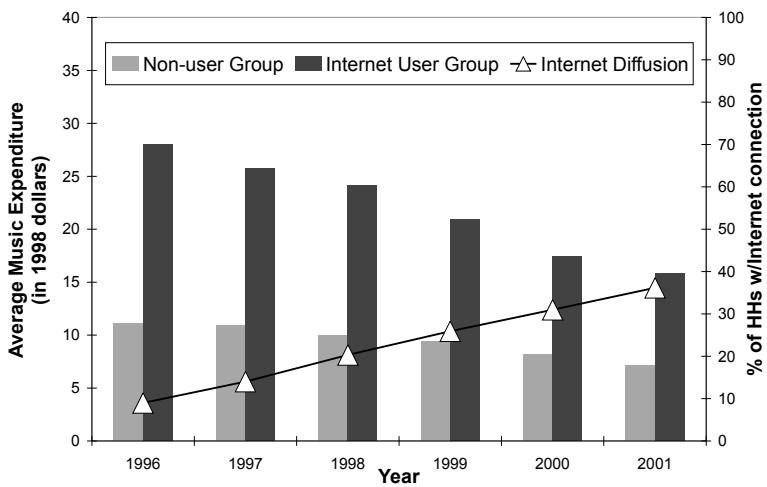


Figure 87: Internet diffusion and music spending

But when we look at Figure 88, we see evidence of compositional changes in the unit itself. While music expenditure fell over the treatment period, the age of the sample grew while income fell. If older people are less likely to buy music in the first place, then this could independently explain some of the decline. This kind of compositional change is a kind of omitted variable bias caused by time-variant unobservables. Diffusion of the Internet appears to be changing the samples as younger music fans are early adopters.

Figure 88: Comparison of Internet user and non-user groups

Table 1: Descriptive Statistics for Internet User and Non-user Groups^a

Year	1997		1998		1999	
	Internet User	Non-user	Internet User	Non-user	Internet User	Non-user
Average Expenditure						
Recorded Music	\$25.73	\$10.90	\$24.18	\$9.97	\$20.92	\$9.37
Entertainment	\$195.03	\$96.71	\$193.38	\$84.92	\$182.42	\$80.19
Zero Expenditure						
Recorded Music	.56	.79	.60	.80	.64	.81
Entertainment	.08	.32	.09	.35	.14	.39
Demographics						
Age	40.2	49.0	42.3	49.0	44.1	49.4
Income	\$52,887	\$30,459	\$51,995	\$28,169	\$49,970	\$26,649
High School Grad.	.18	.31	.17	.32	.21	.32
Some College	.37	.28	.35	.27	.34	.27
College Grad.	.43	.21	.45	.21	.42	.20
Manager	.16	.08	.16	.08	.14	.08

Stata exercise: Abortion legalization and longrun gonorrhea incidence

Exposition of Cunningham and Cornwell [2013] As we have shown, estimating the DD model is straightforward, but running through an example would probably still be beneficial. But since the DDD requires reshaping the data, it would definitely be useful to run through an example that did both. The study we will be replicating is Cunningham and Cornwell [2013]. But first let's learn about the project and the background. Gruber et al. [1999] started a controversial literature. What was the effect that abortion legalization in the 1970s had on the marginal child who would've been born 15-20 years later? The authors showed that the child who would have been born had abortion remained illegal was 60% more likely to live in a single-parent household.

The most famous paper to pick up on that basic stylized fact was Donohue and Levitt [2001]. The authors link abortion legalization in the early 1970s with the decline in crime in the 1990s. Their argument was similar to Gruber et al. [1999] - the marginal child was unwanted and would've grown up in poverty, both of which they argued could predict higher criminal propensity to commit crime as the cohort

aged throughout the age-crime profile. But, abortion legalization (in Gruber et al. [1999] and Donohue and Levitt [2001]'s argument) removed these individuals and as such the treated cohort had positive selection. Levitt [2004] attributes as much as 10% of the decline in crime between 1991 and 2001 to abortion legalization in the 1970s.

This literature was, not surprisingly, incredibly controversial, some of it unwarranted. When asked whether abortion was correct to be legalized, Levitt hedged and said those sorts of ethical questions were beyond the scope of his study. Rather, his was a *positive* study interested only in cause and effect. But some of the ensuing criticism was more legitimate. Joyce [2004], Joyce [2009], and Foote and Goetz [2008] all disputed the findings – some through replication exercises using different data and different identification strategies, and some through the discovery of key coding errors. Furthermore, why look at only crime? If the effect was as large as the authors claim, then wouldn't we find effects *everywhere*?

Cunningham and Cornwell [2013] sought to build on Joyce [2009]'s challenge - if the abortion-selection hypothesis has merit, then shouldn't we find it elsewhere? Because of my research agenda in risky sexual behavior, I chose to investigate the effect on gonorrhea incidence. Why STIs? For one, the characteristics of the marginal child could explain risky sexual behavior that leads to disease transmission. Being raised a single parent is a strong predictor of earlier sexual activity and unprotected sex. Levine et al. [1999] found that abortion legalization caused teen childbearing to fall by 12%. Charles and Luoh [2006] reported that children exposed *in utero* to a legalized abortion regime were less likely to use illegal substances which is correlated with risky sexual behavior.

The estimating strategy that I used was conventional at the time. Five states repealed abortion laws three years before *Roe v. Wade*. My data from the CDC comes in five-year age categories (e.g., 15-19, 20-24 year olds). This created some challenges. First, the early repeal of some states should show declines in gonorrhea for the treated cohort three years before *Roe* states (i.e., the rest of the country). Specifically, we should see lower incidence among 15-19 year olds in the repeal states during the 1986-1992 period relative to their *Roe* counterparts. Second, the treatment effect should be nonlinear because treated cohorts in the repeal states do not fully come of age until 1988, just when the 15-year-olds born under *Roe* enter the sample. Thus we should find negative effects on gonorrhea incidence *briefly* lasting only for the duration of time until *Roe* cohorts catch up and erase the effect. I present a diagram of this dynamic in Figure 89. The top horizontal axis shows the year of the panel; the vertical axis shows the age in calendar years. The cells show the cohort for

those individuals who are of a certain age in that given year. So for instance, a 15-year-old in 1985 was born in 1970. A 15-year-old in 1986 was born in 1971, and so forth. The highlighted blue means that person was exposed to repeal, and the highlighted yellow means that *Roe* catches up.

	CDC Surveillance Data in Calendar Year															
Age in calendar year	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
15	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85
16	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84
17	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83
18	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82
19	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81
20	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80
21	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79
22	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78
23	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77
24	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76
25	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75
26	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74
27	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73
28	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72
29	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71
Number of cohorts (age 15-19) exposed, reforms in 71, 74																
	Repeal (1)	0	1	2	3	4	5	5	5	5	5	5	5	5	5	5
	No Repeal (2)	0	0	0	0	1	2	3	4	5	5	5	5	5	5	5
	Difference (3)	0	1	2	3	3	3	2	1	0	0	0	0	0	0	0

Figure 89: Theoretical predictions of abortion legalization on age profiles of gonorrhea incidence

This creates a very specific age pattern in the treatment effect, represented by the colored bottom row. We should see no effect in 1985; a slightly negative effect in 1986 as the first cohort reaches 15, an even more negative effect through 1987 and an even more negative effect from 1988-1990. But then following 1990 through 1992, the treatment effect should gradually disappear. All subsequent DD coefficients should be zero thereafter since there is no difference at that point in the *Roe* and repeal states beyond 1992.

A simple graphic for Black 15-19 year old female incidence can help illustrate our findings. Remember, a picture speaks a thousand words, and whether it's RDD or DD, it's helpful to show pictures like these to prepare the reader for the table after table of regression coefficients. I present two pictures; one showing the raw data, and one showing the DD coefficients. The first is Figure 91. This picture captures the dynamics that we will be picking up in our DD plots. The shaded areas represent the period of time where differences between the treatment and control units should be different, and beyond they should be the same, conditional on a state and year fixed effect. And as you can see, *Roe* states experienced a large increase in gonorrhea

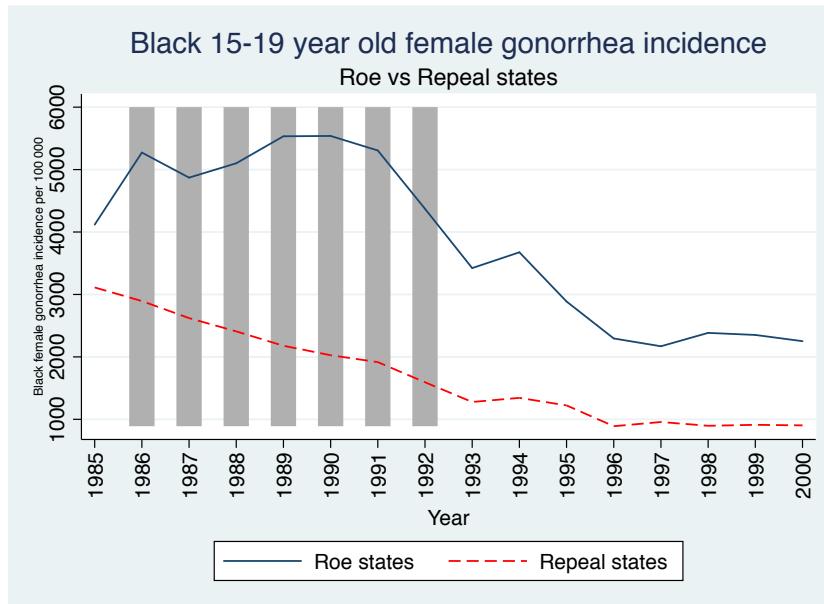


Figure 90: Differences in black female gonorrhea incidence between repeal and *Roe* cohorts.

during the window where repeal states were falling.

Our estimating equation is as follows:

$$Y_{st} = \beta_1 Repeals + \beta_2 DT_t + \beta_3 Repeals \times DT_t + X_{st}\psi + \alpha_s DS_s + \gamma_1 t + \gamma_2 s \times t + \varepsilon_{st}$$

where Y is the log number of new gonorrhea cases for 15-19 year olds (per 100,000 of the population); $Repeals$ equals one if the state legalized abortion prior to *Roe*; DT_t is a year dummy; DS_s is a state dummy; t is a time trend; X is a matrix of covariates; $DS_s \times t$ are state specific linear trends; and ε_{st} is an error term assumed to be conditionally independent of the regressors. We present plotted coefficients from this regression for simplicity (and because pictures can be so powerful) in Figure 91. As can be seen, there is a negative effect during the window where *Roe* has not fully caught up:

The regression equation for a DDD is more complicated as you recall from the Gruber [1994] paper. Specifically, it requires stacking new comparison within-state units who capture state-specific trends but who were technically untreated. We chose the 25-29 year olds in the same states as within-state comparison groups. We also chose the 20-24 year olds as a within-state comparison group but our reasoning was that that age group, while not treated, was more likely to have sex with the 15-19 year olds, who were treated, and thus SUTVA was violated. So we chose a group that was reasonably close to capture trends, but not so close that they violate SUTVA. The estimating

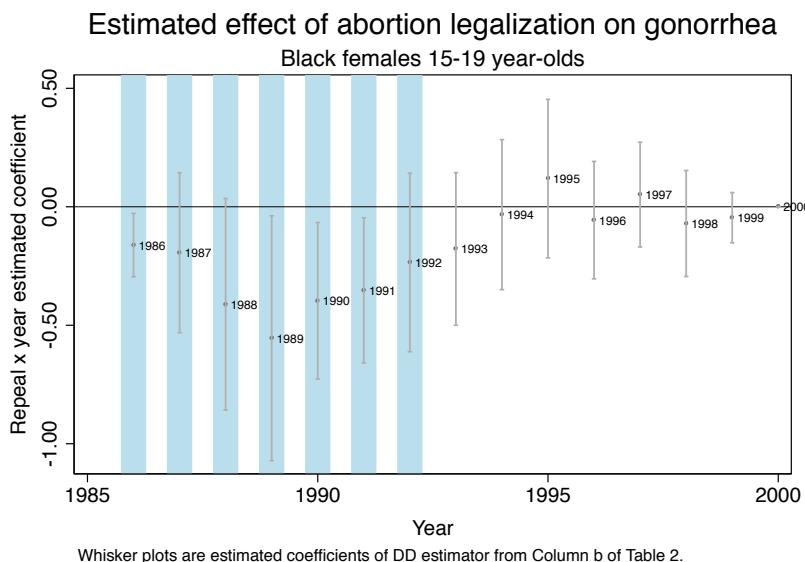


Figure 91: Coefficients and standard errors from DD regression equation

equation for this regression is

$$\begin{aligned}
 Y_{ast} = & \beta_1 Repeal_s + \beta_2 DT_t + \beta_{3t} Repeal_s \cdot DT_t + \delta_1 DA + \delta_2 Repeal_s \cdot DA \\
 & + \delta_{3t} DA \cdot DT_t + \delta_{4t} Repeal_s \cdot DA \cdot DT_t + X_{st}\xi + \alpha_{1s} DS_s + \alpha_{2s} DS_s \cdot DA \\
 & + \gamma_1 t + \gamma_2 DS_s \cdot t + \gamma_3 DA \cdot t + \gamma_4 DS_s \cdot DA \cdot t + \epsilon_{ast},
 \end{aligned}$$

where the DDD parameter we are estimating is γ_4 - the full interaction. In case this wasn't obvious, the reason there are 8 separate dummies is because our DDD parameter has all three interactions. Thus since there are 9 combinations, we had to drop one as the omitted group, and control separately for the other 7. Here we present the table of coefficients. Note that the effect should be concentrated only among the treatment years as before. This is presented here in Figure 92:¹³⁶ Column (b) controls for an age-state interaction with age-state specific linear time trends. As can be seen, we see nearly the same pattern using DDD as we found with our DD, though the precision is smaller. I interpreted these patterns as evidence for the original Gruber et al. [1999] and Donohue and Levitt [2001] abortion-selection hypothesis.

Stata replication Now what I'd like to do is replicate some of these results, as I want you to have handy a file that will estimate a DD model, but also the slightly more cumbersome DDD model. Before we begin, you will need to download `cgmreg.ado` from Doug Miller's website, as referees asked us to implement the multi-way clustering correction for the standard errors to allow for correlation both across

¹³⁶ Note, because the original table spans multiple pages, I didn't want to clutter up the page with awkwardly linked tables. But you can see the full table on pages 401-402 of Cunningham and Cornwell [2013].

Covariates	Black female	
	(a)	(b)
Repeal \times 15-year-old \times 1986	-0.337*** (0.115)	-0.389*** (0.126)
Repeal \times 15-year-old \times 1987	-0.389** (0.155)	-0.451** (0.189)
Repeal \times 15-year-old \times 1988	-0.382** (0.143)	-0.472** (0.182)
Repeal \times 15-year-old \times 1989	-0.277* (0.138)	-0.380* (0.191)
Repeal \times 15-year-old \times 1990	-0.046 (0.146)	-0.163 (0.169)
Repeal \times 15-year-old \times 1991	0.079 (0.148)	-0.039 (0.216)
Repeal \times 15-year-old \times 1992	0.122 (0.140)	0.005 (0.144)
Repeal \times 15-year-old \times 1993	-0.168 (0.360)	-0.261 (0.328)
Repeal \times 15-year-old \times 1994	0.239* (0.124)	0.112 (0.104)
Repeal \times 15-year-old \times 1995	0.151 (0.142)	0.060 (0.096)
Repeal \times 15-year-old \times 1996	0.183 (0.114)	0.095 (0.115)
Repeal \times 15-year-old \times 1997	0.357*** (0.114)	0.269*** (0.098)

Figure 92: Subset of coefficients (year-repeal interactions) for the DDD model, Table 3 of Cunningham and Cornwell [2013].

states and within states. That can be found at the top of <http://faculty.econ.ucdavis.edu/faculty/dlmiller/Statafiles/>, and as with scuse.ado, must simply be saved into the /c subdirectory of your Stata folders. Let's begin:

```
. scuse abortion, clear
. xi: cgmreg lnr i.repeal*i.year i.fip acc ir pi alcohol crack poverty income ur if bf15==1 //
[aweight=totpop], cluster(fip year)
. test _IrepXyea_1_1986 _IrepXyea_1_1987 _IrepXyea_1_1988 _IrepXyea_1_1989 //
_IrepXyea_1_1990 _IrepXyea_1_1991 _IrepXyea_1_1992
```

The last line tests for the joint significance of the treatment (repeal \times year interactions). Note, for simplicity, I only estimated this for the black females (`bf15==1`) but you could estimate for the black males (`bm15==1`), white females (`wf15==1`) or white males (`wm15==1`). We do all four in the paper, but I am just trying to give you a basic understanding of the syntax.

Next, we show how to use this sample so that we can estimate a DDD model. A considerable amount of reshaping had to be done earlier in the code, but it would take too long to post that here, so in v. 2.0 of this book, I will provide the do file that was used to make the tables for this paper. For now, though, I will simply produce the commands that produce the black female result.

```

. gen yr=(repeal==1) & (younger==1)
. gen wm=(wht==1) & (male==1)
. gen wf=(wht==1) & (male==0)
. gen bm=(wht==0) & (male==1)
. gen bf=(wht==0) & (male==0)
. char year[omit] 1985
. char repeal[omit] 0
. char younger[omit] 0
. char fip[omit] 1
. char fa[omit] 0
. char yr[omit] 0
.xi: cgmreg lnr i.repeal*i.year i.younger*i.repeal i.younger*i.year i.yr*i.year //
    i.fip*t acc pi ir alcohol crack poverty income ur if 'x'==1 & (age==15 | age==25) //
    [aweight=totpop], cluster(fip year)
. test _IyrXyea_1_1986 _IyrXyea_1_1987 _IyrXyea_1_1988 _IyrXyea_1_1989 _IyrXyea_1_1990 //
    _IyrXyea_1_1991 _IyrXyea_1_1992

```

Notice that some of these already are interactions (e.g., `yr`) which was my way to compactly include all of the interactions since at the time my workflow used the asterisk to create interactions as opposed to the hashtag (e.g., `#`). But I encourage you to study the data structure itself. Notice how I used if-statements to limit the regression analysis which forced the data structure to shrink into either the DD matrix or the DDD matrix depending on how I did it.

Conclusion I have a bumper sticker on my car that says “I love Federalism (for the natural experiments)” (Figure 93).¹³⁷ The reason



¹³⁷ Although this one has a misspelling, the real one has the plural version of experiments. I just couldn't find the Figure 93: I \heartsuit Federalism bumper-sticker (for the natural experiments)

I made this was half tongue-in-cheek, half legitimate gratitude. Because of state federalism, each American state is allowed considerable legislative flexibility to decide its own governance and laws. Yet,

because of the federal government, many of our datasets are harmonized across states, making it even more useful for causal inference than European countries which do not always have harmonized datasets for many interesting questions outside of macroeconomics.

The reason to be grateful for federalism is that it provides a constantly evolving laboratory for applied researchers seeking to evaluate the causal effects of laws and other interventions. It has therefore for this reason probably become one of the most popular forms of identification among American researchers, if not the most common. A google search of the phrase “differences in differences” brought up 12 million hits. It is arguably the most common methodology you will use – moreso than IV or matching or even RDD, despite RDD’s greater credibility. There is simply a never ending flow of quasi-experiments being created by our decentralized data generating process in the United States made even more advantageous by so many federal agencies being responsible for data collection, thus ensuring improved data quality and consistency.

Study the ideas in this chapter. Review them. Review the dataset I provided and the Stata syntax. Walk yourself through the table and figures I presented. Think carefully about why the regression analysis reproduces the exact same differencing that we presented in our DD and DDD tables. Study the DAG at the start of the chapter and the formal technical assumptions necessary for identification. Understanding what you’re doing in DD and DDD is key to your career because of its popularity if nothing else. You need to understand how it works, and under what conditions it can identify causal effects, if only to interact with colleagues and peers’ research.

Synthetic control

"The synthetic control approach developed by [Abadie et al. \[2010, 2015\]](#) and [Abadie and Gardeazabal \[2003\]](#) is arguably the most important innovation in the policy evaluation literature in the last 15 years." - [Athey and Imbens \[2017\]](#)

In qualitative case studies, such as de Tocqueville's classic [Democracy in America](#), the goal is to reason inductively about the causal effect of events or characteristics of a single unit on some outcome using logic and historical analysis. But it may not give a very satisfactory answer to these causal questions because oftentimes it lacks a counterfactual. As such, we are usually left with description and speculation about the causal pathways connecting various events to outcomes.

Quantitative comparative case studies are more explicitly causal designs. They usually are natural experiments and they usually are applied to only a single unit, such as a single school, firm, state or country. These kinds of quantitative comparative case studies compare the evolution of an aggregate outcome with either some single other outcome, or as is more oftentimes the case, a chosen set of similar units which serve as a control group.

As [Athey and Imbens \[2017\]](#) point out, one of the most important contributions to quantitative comparative case studies is the synthetic control model. The synthetic control model was developed in [Abadie and Gardeazabal \[2003\]](#) in a study of terrorism's effect on aggregate income which was then elaborated on in a more exhaustive treatment [\[Abadie et al., 2010\]](#). Synthetic controls models optimally choose a set of weights which when applied to a group of corresponding units produce an optimally estimated counterfactual to the unit that received the treatment. This counterfactual, called the "synthetic unit", serves to outline what would have happened to the aggregate treated unit had the treatment never occurred. It is a powerful, yet surprisingly simple, generalization of the differences-in-differences strategy. We will discuss it now with a motivating example - the famous Mariel boatlift paper by [Card \[1990\]](#).

Cuba, Miami and the Mariel Boatlift

Labor economists have debated the effect of immigration on local labor market conditions for many years [Card and Peri, 2016]. Do inflows of immigrants depress wages and the employment of natives in local labor markets? For Card [1990], this was an empirical question, and he used a natural experiment to evaluate it.

In 1980, Fidel Castro announced that anyone wishing to leave Cuba could do so if they exited from Mariel by a certain date, called the Mariel Boatlift. The Mariel Boatlift was a mass exodus from Cuba's Mariel Harbor to the United States (primarily Miami Florida) between April and October 1980. Approximately 125,000 Cubans emigrated to Florida over this six month period of time. The emigration stopped only because Cuba and the US mutually agreed to end it. The event increased the Miami labor force by 7%, largely by depositing a record number of low skill workers into a relatively small area.

Card saw this as an ideal natural experiment. It was arguably an exogenous shift in the labor supply curve, which would allow him to determine if wages fell and employment increased, consistent with a simple competitive labor market model. He used individual-level data on unemployment from the CPS for Miami and chose four comparison cities (Atlanta, Los Angeles, Houston and Tampa-St. Petersburg). The choice of these four cities is delegated to a footnote in the paper wherein Card argues that they were similar based on demographics and economic conditions. Card estimated a simple DD model and found, surprisingly, no effect on wages or native unemployment. He argued that Miami's labor market was capable of absorbing the surge in labor supply because of similar surges two decades earlier.

The paper was very controversial, probably not so much because he attempted to answer empirically an important question in labor economics using a natural experiment, but rather because the result violated conventional wisdom. It would not be the last word on the subject, and I don't take a stand on this question; rather, I introduce it to highlight a few characteristics of the study.

It was a comparative case study which had certain strengths and weaknesses. The policy intervention occurred at an aggregate level, for which aggregate data was available. But the problems with the study were that the selection of the control group is ad hoc and ambiguous, and secondly, the standard errors reflect sampling variance as opposed to uncertainty about the ability of the control group to reproduce the counterfactual of interest.¹³⁸

¹³⁸ Interestingly, a recent study replicated Card's paper using synthetic control and found similar results. [Peri and Yasenov, 2018].

Abadie and Gardeazabal [2003] and Abadie et al. [2010] introduced the synthetic control estimator as a way of addressing both simultaneously. This method uses a weighted average of units in the donor pool to model the counterfactual. The method is based on the observation that, when the units of analysis are a few aggregate units, a combination of comparison units (the “synthetic control”) often does a better job of reproducing characteristics of a treated unit than using a single comparison unit alone. The comparison unit, therefore, in this method is selected to be the weighted average of all comparison units that best resemble the characteristics of the treated unit(s) in the pre-treatment period.

Abadie et al. [2010] argue that this method has many distinct advantages over regression based methods. For one, the method precludes extrapolation. It uses instead interpolation, because the estimated causal effect is always based on a comparison between some outcome in a given year and a counterfactual in the same year. That is, its uses as its counterfactual a convex hull of control group units, and thus the counterfactual is based on where data actually is, as opposed to extrapolating beyond the support of the data which can occur in extreme situations with regression [King and Zeng, 2006].

A second advantage has to do with processing of the data. The construction of the counterfactual does not require access to the post-treatment outcomes during the design phase of the study, unlike regression. The advantage here is that it helps the researcher avoid “peaking” at the results while specifying the model. Care and honesty must still be used, as it’s just as easy to also look at the outcomes during the design phase as it is to not, but the point is that it is hypothetically possible to focus just on design, and not estimation, with this method.

Another advantage, which is oftentimes a reason that people will object to the study ironically, is that the weights which are chosen make explicit what each unit is contributing the counterfactual. Now this is in many ways a strict advantage, except when it comes to defending those weights in a seminar. Because someone can see that Idaho is contributing 0.3 to your modeling of Florida, they are now able to argue that it’s absurd to think Idaho is anything like Florida. But contrast this with regression, which also weights the data, but does so blindly. The only reason no one objects to what regression produces as a weight is that they *cannot see the weights*. They are implicit, rather than explicit. So I see this explicit production of weights as a distinct advantage because it makes synthetic control more transparent than regression based designs.

A fourth advantage, which I think is often unappreciated, is

that it bridges a gap between qualitative and quantitative types. Qualitative researchers are often the very ones focused on describing a single unit, such as a country or a prison [Perkinson, 2010], in great detail. They are usually the experts on the histories surrounding those institutions. They are usually the ones doing comparative case studies in the first place. Synthetic control places a valuable tool into their hands which enables them to choose counterfactuals - a process that in principle can improve their work insofar as they are interested in evaluating some particular intervention.

Finally, Abadie et al. [2010] argue that it removes subjective researcher bias, but I actually believe this is the most overstated benefit of the method. Through repeated iterations and changes to the matching formula, a person can just as easily introduce subjective choices into the process. Sure, the weights are optimally chosen to minimize some distance function, but through the choice of the covariates themselves, the researcher can in principle select different weights. She just doesn't have a lot of control over it, because ultimately the weights are optimal for a given set of covariates.

Formalization Let Y_{jt} be the outcome of interest for unit j of $J + 1$ aggregate units at time t , and treatment group be $j = 1$. The synthetic control estimator models the effect of the intervention at time T_0 on the treatment group using a linear combination of optimally chosen units as a synthetic control. For the post-intervention period, the synthetic control estimator measures the causal effect as $Y_{1t} - \sum_{j=2}^{J+1} w_j^* Y_{jt}$ where w_j^* is a vector of optimally chosen weights.

Matching variables, X_1 and X_0 , are chosen as predictors of post-intervention outcomes and must be unaffected by the intervention. The weights are chosen so as to minimize the norm, $\|X_1 - X_0 W\|$ subject to weight constraints. There are two weight constraints. First, let $W = (w_2, \dots, w_{J+1})'$ with $w_j \geq 0$ for $j = 2, \dots, J + 1$. Second, let $w_2 + \dots + w_{J+1} = 1$. In words, no unit receives a negative weight, but can receive a zero weight.¹³⁹ And the sum of all weights must equal one.

As I said, Abadie et al. [2010] consider

$$\|X_1 - X_0 W\| = \sqrt{(X_1 - X_0 W)' V (X_1 - X_0 W)}$$

where V is some $(k \times k)$ symmetric and positive semidefinite matrix. Let X_{jm} be the value of the m -th covariates for unit j . Typically, V is diagonal with main diagonal v_1, \dots, v_k . Then the synthetic control weights minimize:

$$\sum_{m=1}^k v_m \left(X_{1m} - \sum_{j=2}^{J+1} w_j X_{jm} \right)^2$$

¹³⁹ See Doudchenko and Imbens [2016] for recent work relaxing the non-negativity constraint.

where v_m is a weight that reflects the relative importance that we assign to the m -th variable when we measure the discrepancy between the treated unit and the synthetic control.

The choice of V , as should be seen by now, is important because W^* depends on one's choice of V . The synthetic control $W^*(V)$ is meant to reproduce the behavior of the outcome variable for the treated unit in the absence of the treatment. Therefore, the weights v_1, \dots, v_k should reflect the predictive value of the covariates.

[Abadie et al. \[2010\]](#) suggests different choices of V , but ultimately it appears from practice that most people choose V that minimizes the mean squared prediction error:

$$\sum_{t=1}^{T_0} \left(Y_{1t} - \sum_{j=2}^{J+1} w_j^*(V) Y_{jt} \right)^2$$

What about unobserved factors? Comparative case studies are complicated by unmeasured factors affecting the outcome of interest as well as heterogeneity in the effect of observed and unobserved factors. [Abadie et al. \[2010\]](#) note that if the number of pre-intervention periods in the data is “large”, then matching on pre-intervention outcomes can allow us to control for the heterogeneous responses to multiple unobserved factors. The intuition here is that only units that are alike on unobservables and unobservables would follow a similar trajectory pre-treatment.

California's Proposition 99 [Abadie and Gardeazabal \[2003\]](#) developed the synthetic control estimator so as to evaluate the impact that terrorism had on the Basque region. But [Abadie et al. \[2010\]](#) expounds on the method by using a cigarette tax in California called Proposition 99. The cigarette tax example uses a placebo-based method for inference, which I'm wanting to explain, so let's look more closely at their paper.

In 1988, California passed comprehensive tobacco control legislation called Proposition 99. Proposition 99 increased cigarette taxes by 25 cents a pack, spurred clean-air ordinances throughout the state, funded anti-smoking media campaigns, earmarked tax revenues to health and anti-smoking budgets, and produced more than \$100 million a year in anti-tobacco projects. Other states had similar control programs, and they were dropped from their analysis.

Figure 94 shows changes in cigarette sales for California and the rest of the United States annually from 1970 to 2000. As can be seen, cigarette sales fell after Proposition 99, but as they were already falling, it's not clear if there was any effect – particularly since they were falling in the rest of the country at the same time.

Using their method, though, they select an optimal set of weights

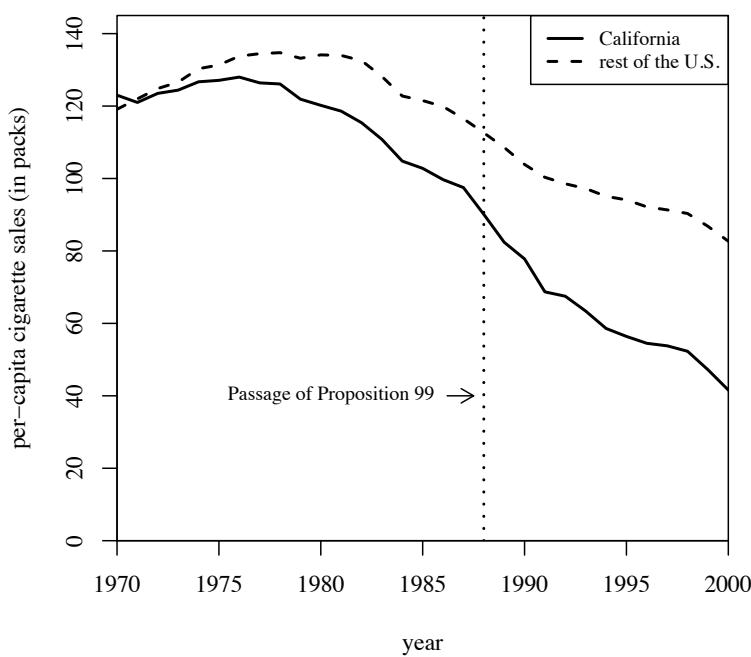


Figure 94: California cigarette sales vs the rest of the country

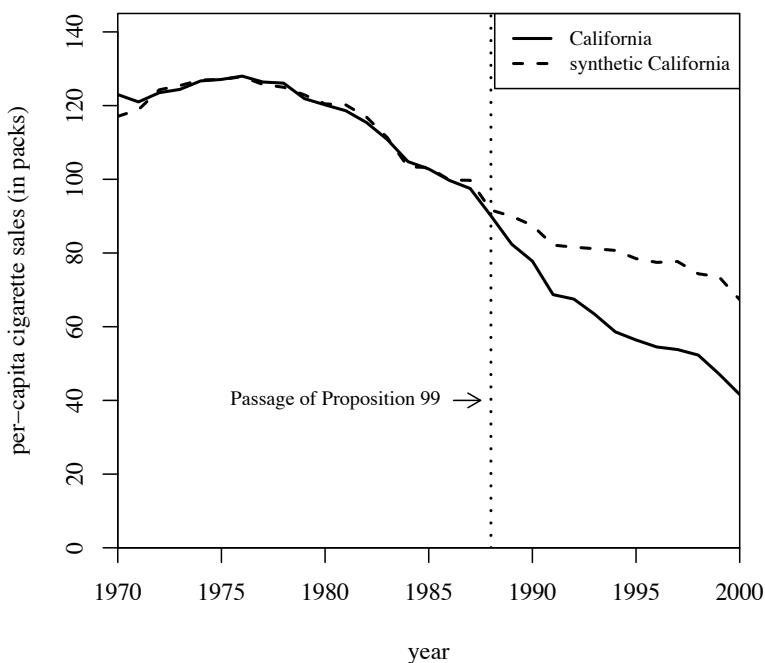


Figure 95: California cigarette sales vs synthetic California

that when applied to the rest of the country produces the figure shown in Figure 95. Notice that pre-treatment, this set of weights produces a nearly identical time path for California as the real California itself, but post-treatment the two series diverge. There appears at first glance to have been an effect of the program on cigarette sales.

The variables they used for their distance minimization are listed in Figure 96. Notice that this analysis produces values for the treatment group and control group that facilitate a simple investigation of balance. This is not a technical test, as there are only one value per variable per treatment category, but it's the best we can do with this method. And it appears that the variables used for matching are similar across the two groups, particularly for the lagged values.

Figure 96: Balance table

Variables	California		Average of
	Real	Synthetic	38 control states
Ln(GDP per capita)	10.08	9.86	9.86
Percent aged 15-24	17.40	17.40	17.29
Retail price	89.42	89.41	87.27
Beer consumption per capita	24.28	24.20	23.75
Cigarette sales per capita 1988	90.10	91.62	114.20
Cigarette sales per capita 1980	120.20	120.43	136.58
Cigarette sales per capita 1975	127.10	126.99	132.81

Note: All variables except lagged cigarette sales are averaged for the 1980–1988 period (beer consumption is averaged 1984–1988).

Like RDD, synthetic control is a picture-intensive estimator. Your estimator is basically a picture of two series which, if there is a causal effect, diverge from another post-treatment, but resemble each other pre-treatment. It is common to therefore see a picture just showing the difference between the two series (Figure 97). But so far, we have only covered estimation. How do we determine whether the observed difference between the two series is a *statistically significant* difference? After all, we only have two observations per year. Maybe the divergence between the two series is nothing more than prediction error, and any model chosen would've done that, even if there was no treatment effect. Abadie et al. [2010] suggest that we use an old fashioned method to construct exact p-values based on Fisher [1935]. This is done through “randomization” of the treatment to each unit, re-estimating the model, and calculating a set of root mean squared prediction error (RMSPE) values for the pre- and post-treatment period.¹⁴⁰ We proceed as follows:

1. Iteratively apply the synthetic control method to each country/state in the donor pool and obtain a distribution of placebo

¹⁴⁰ What we will do is simply reassign the treatment to each unit, putting California back into the donor pool each time, estimate the model for that “placebo”, and recording information from each iteration.

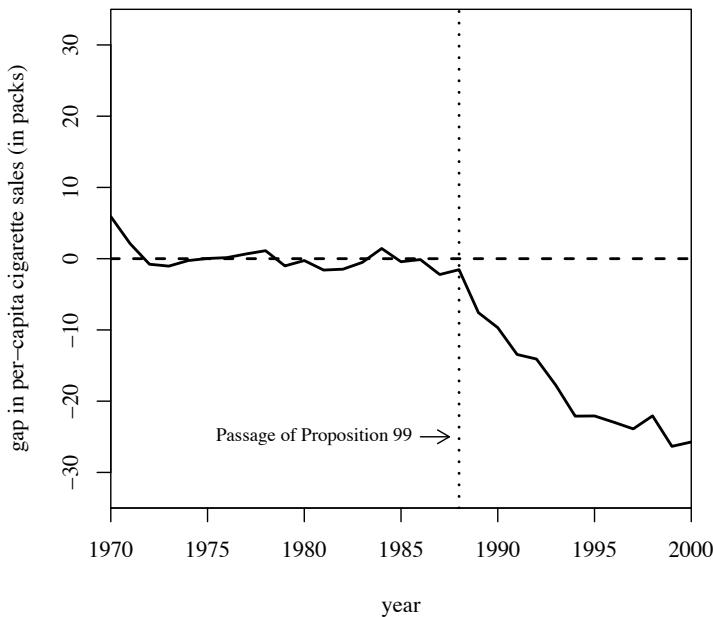


Figure 97: California cigarette sales vs synthetic California

effects

2. Calculate the RMSPE for each placebo for the pre-treatment period:

$$RMSPE = \left(\frac{1}{T - T_0} \sum_{t=T_0+t}^T \left(Y_{1t} - \sum_{j=2}^{J+1} w_j^* Y_{jt} \right)^2 \right)^{\frac{1}{2}}$$

3. Calculate the RMSPE for each placebo for the post-treatment period (similar equation but for the post-treatment period)
4. Compute the ratio of the post-to-pre-treatment RMSPE
5. Sort this ratio in descending order from greatest to highest.
6. Calculate the treatment unit's ratio in the distribution as $p = \frac{\text{RANK}}{\text{TOTAL}}$

In other words, what we want to know is whether California's treatment effect is extreme, which is a relative concept compared to the donor pool's own placebo ratios.

There's several different ways to represent this. The first is to overlay California with all the placebos using Stata `twoway` command, which I'll show later. Figure 98 shows what this looks like. And I think you'll agree, it tells a nice story. Clearly, California is in the tails of some distribution of treatment effects. [Abadie et al. \[2010\]](#) recommend iteratively dropping the states whose pre-treatment

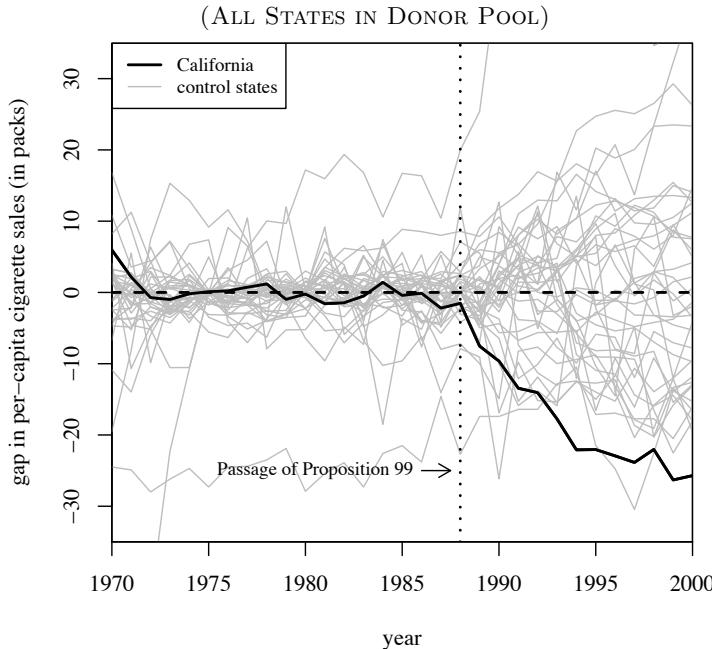


Figure 98: Placebo distribution

RMSPE is considerably different than California's because as you can see, they're kind of blowing up the scale and making it hard to see what's going on. They do this in several steps, but I'll just skip to the last step (Figure 99). In this, they've dropped any state unit from the graph whose pre-treatment RMSPE is more than two times that of California's. This therefore limits the picture to just units whose model fit, pre-treatment, was pretty good, like California's. But, ultimately, inference is based on those exact p-values. So the way we do this is we simply create a histogram of the ratios, and more or less mark the treatment group in the distribution so that the reader can see the exact p-value associated with the model. I produce that here in Figure 100. As can be seen, California is ranked 1st out of 38 state units.¹⁴¹ This gives an exact p-value of 0.026, which is less than the conventional 5% most journals want to (arbitrarily) see for statistical significance.

Falsifications In Abadie et al. [2015], the authors studied the effect of the reunification of Germany on GDP. One of the contributions this paper makes, though, is a recommendation for how to test the validity of the estimator through a falsification exercise. To illustrate this, let's walk through their basic findings. In Figure 101, the authors illustrate their main question by showing the changing trend lines for West Germany and the rest of their OECD sample.

¹⁴¹ Recall, they dropped several states who had similar legislation passed over this time period.

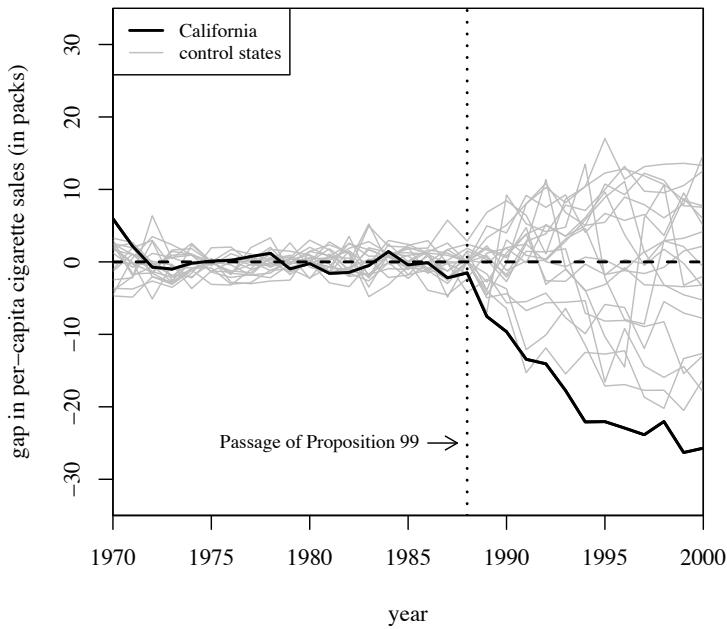
(PRE-PROP. 99 MSPE \leq 2 TIMES PRE-PROP. 99 MSPE FOR CA)

Figure 99: Placebo distribution

(ALL 38 STATES IN DONOR POOL)

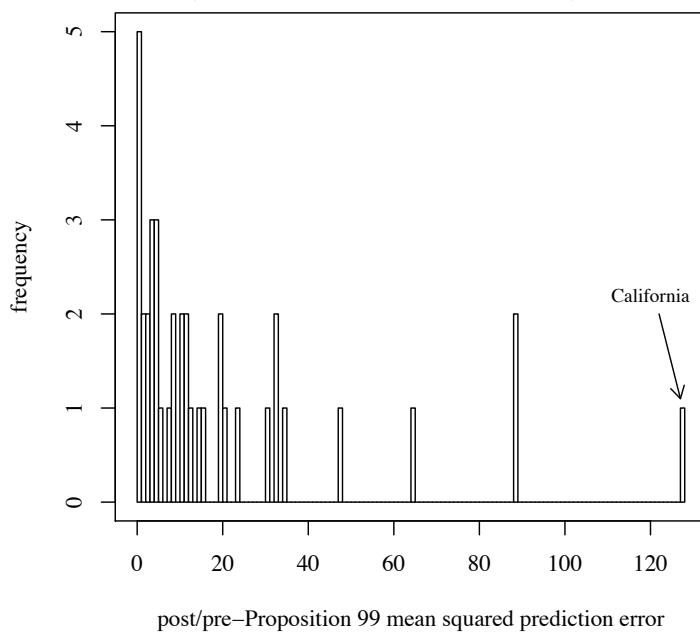
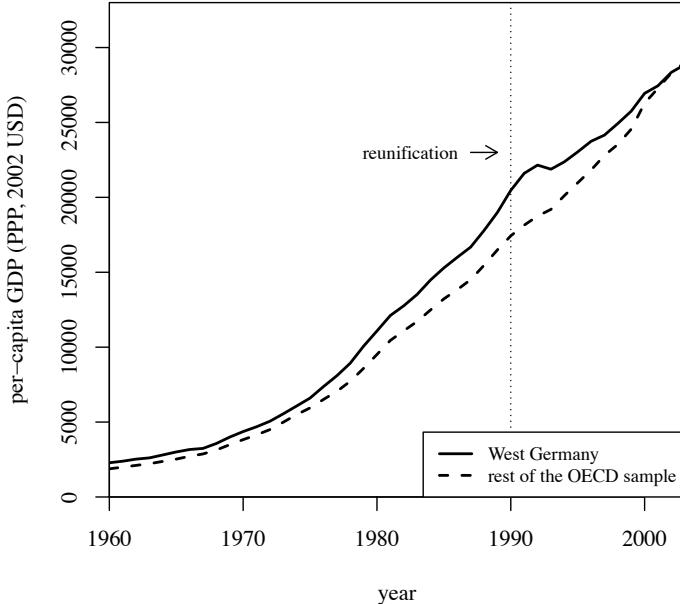


Figure 100: Placebo distribution

Figure 1: Trends in Per-Capita GDP: West Germany vs. Rest of OECD Sample

Figure 101: West Germany GDP vs. Other Countries



As we saw with cigarette smoking, it's difficult to make a statement about the effect of reunification given West Germany is dissimilar from the other countries on average before reunification.

In Figure 101 and Figure 103, we see their main results. The authors then implement the placebo-based inference to calculate exact p -values and find that the estimated treatment effect from reunification is statistically significant.

The placebo-based inference suggests even further robustness checks, though. The authors specifically recommend rewinding time from the date of the treatment itself and estimating their model on an earlier (placebo) date. There should be no effect when they do this; if there is, then it calls into question the research design. The authors do this in Figure 104. Notice that when they run their model on the placebo date of 1975, they ultimately find no effect. This suggests that their model has good in and out of sample predictive properties. Hence since the model does such a good job of predicting GDP per capita, the fact that it fails to anticipate the change in the year of reunification suggests that the model was picking up a causal effect.

We include this second paper primarily to illustrate that synthetic control methods are increasingly expected to pursue numerous

Figure 2: Trends in Per-Capita GDP: West Germany vs. Synthetic West Germany

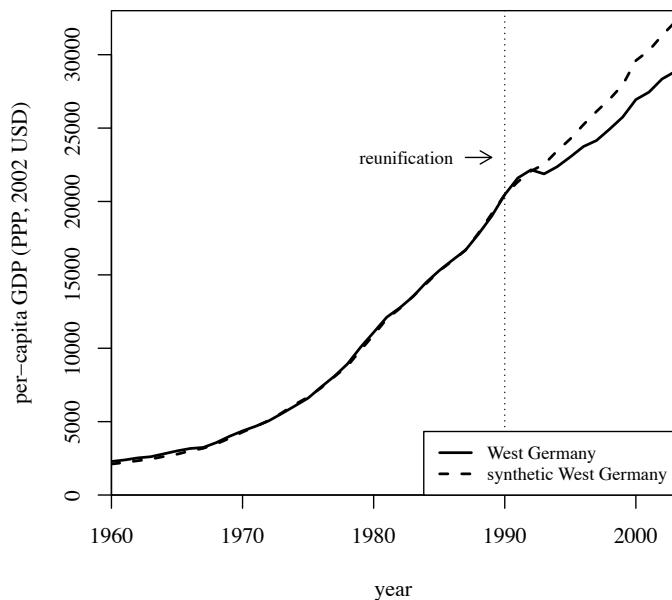


Figure 3: Per-Capita GDP Gap Between West Germany and Synthetic West Germany

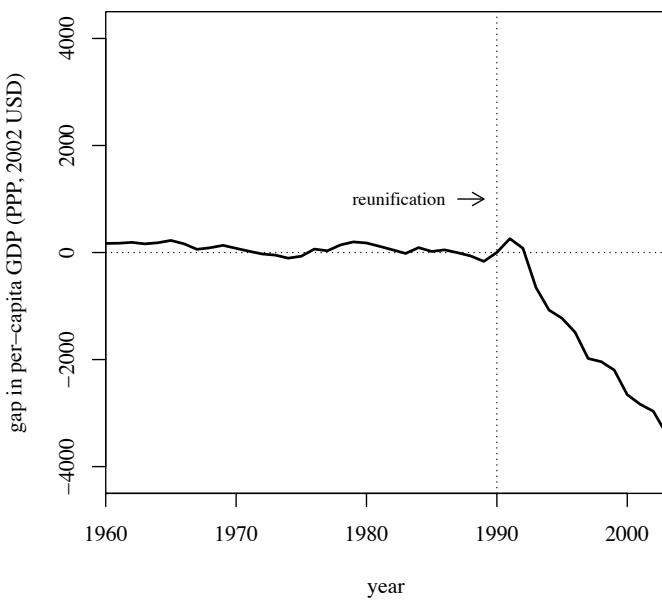


Figure 102: Synthetic control graph: West Germany vs Synthetic West Germany

Figure 103: Synthetic control graph: Differences between West Germany and Synthetic West Germany

Figure 4: Placebo Reunification 1975 - Trends in Per-Capita GDP: West Germany vs. Synthetic West Germany

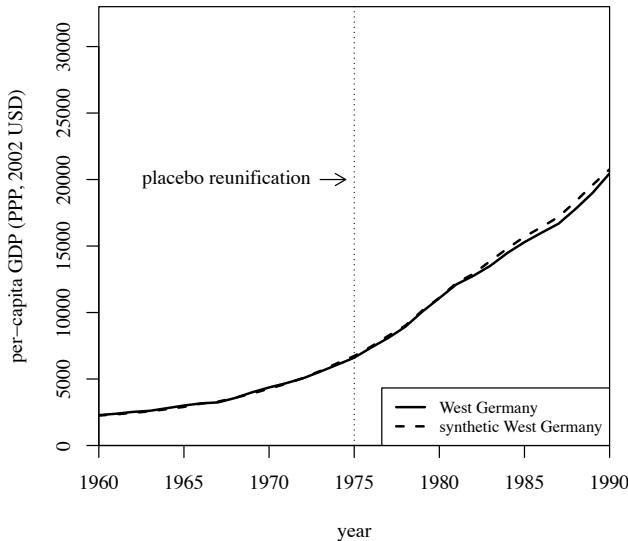


Figure 104: Synthetic control graph: Placebo Date

falsification exercises in addition to simply estimating the causal effect itself. In this sense, researchers have pushed others to hold it to the same level of scrutiny and skepticism as they have with other methodologies such as RDD and IV. Authors using synthetic control must do more than merely run the synth command when doing comparative case studies. They must find the exact p -values through placebo-based inference, check for the quality of the pre-treatment fit, investigate the balance of the covariates used for matching, and check for the validity of the model through placebo estimation (e.g., rolling back the treatment date).

Stata exercise: Prison construction and Black male incarceration

The project that you'll be replicating here is a project I have been working on with several coauthors over the last few years.¹⁴² Here's the backdrop.

In 1980, Texas Department of Corrections lost a major civil action lawsuit. The lawsuit was called *Ruiz v. Estelle*; Ruiz was the prisoner who brought the case, and Estelle was the warden. The case argued that TDC was engaging in unconstitutional practices related to overcrowding and other prison conditions. Surprisingly, Texas lost

¹⁴² You can find one example of an unpublished manuscript here coauthored with Sam Kang: http://scunning.com/prison_booms_and_drugs_20.pdf

the case, and as a result, Texas was forced to enter into a series of settlements. To amend the issue of overcrowding, the courts placed constraints on the number of housing inmates that could be placed in cells. To ensure compliance, TDC was put under court supervision until 2003.

Given these constraints, the construction of new prisons was the only way that Texas could adequately meet demand without letting prisoners go, and since the building of new prisons was erratic, the only other option was increasing the state's parole rate. That is precisely what happened; following *Ruiz v. Estelle*, Texas used paroles more intensively to handle the increased arrest and imprisonment flows since they did not have the operational capacity to handle that flow otherwise.

But, then the state began building prisons which started somewhat in the late 1980s under Governor Bill Clements. However, the prison construction under Clements was relatively modest. Not so in 1993 when Governor Ann Richards embarked on a major prison construction drive. Under Richards, state legislators approved a billion dollar prison construction project which doubled the state's operational capacity within 3 years. This can be seen in Figure 105. As can be

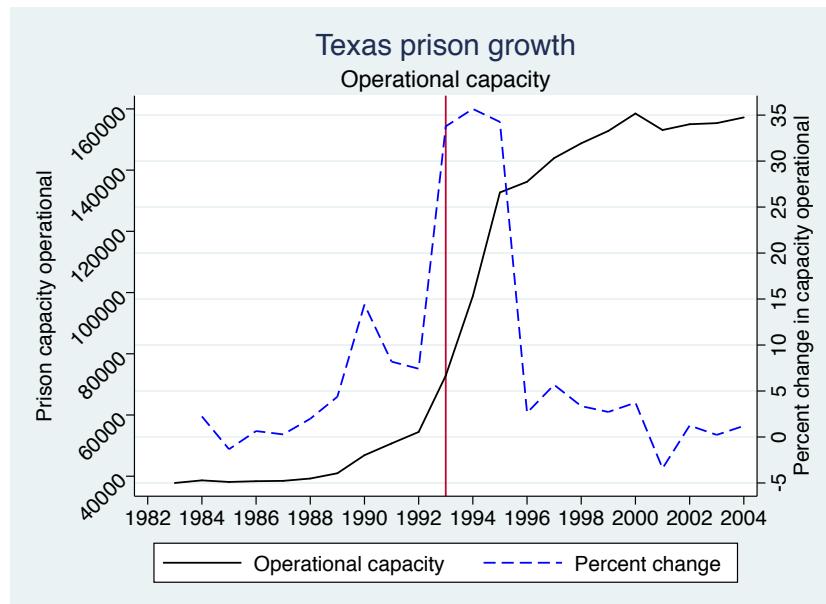


Figure 105: Prison capacity (operational capacity) expansion

seen, Clements build out was relatively modest both as a percentage change and in levels. But Richards' investments in operational capacity was gigantic – the number of beds grew over 30% for three years causing the number of beds to more than double in a short period of time.

What was the effect of building so many prisons? Just because prison capacity expands doesn't mean incarceration rates will grow. But because the state was intensively using paroles to handle the flow, that's precisely what did happen. Because our analysis in a moment will focus on African-American male imprisonment, I will show the effect of the prison boom on African-American male incarceration. As you can see from Figure 106, the Black male incarceration rate

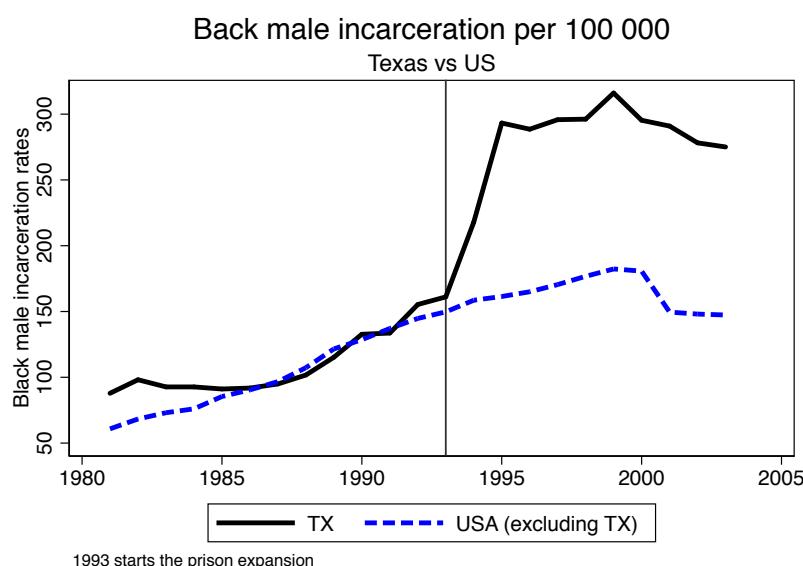


Figure 106: African-American male incarceration rates

went from 150 to 350 in only two years. Texas basically went from being a typical, modal state when it came to incarceration rates to one of the most severe in only a few short periods of time.

What we will now do is analyze the effect that the prison construction under Richards had on Black male incarceration rates using synthetic control. The do file to do this can be downloaded directly from my website at <http://scunning.com/texas.do>, and I probably would recommend downloading it now instead of using the code I'm going to post here. But, let's start now. You'll first want to look at the readme document to learn how to organize a set of subdirectories, as I use subdirectories extensively in this do file. That readme can be found at <http://scunning.com/readme-2.pdf>. The subdirectories you'll need are the following:

- Do
- Data
 - synth

- Inference
- Figures

And I recommend having a designated main directory for all this, perhaps /Texas. In other words the Do directory would be located in /Texas/Do. Now let's begin.

The first step is to create the figure showing the effect of the 1993 prison construction on Black male incarceration rates. I've chosen a set of covariates and pre-treatment outcome variables for the matching; I encourage you, though, to play around with different models. We can already see, though, from Figure 106 that prior to 1993, Texas Black male incarceration rates were pretty similar to the rest of the country. What this is going to mean for our analysis is that we have every reason to believe that the convex hull likely exists in this application.

```
.cd "/users/scott_cunningham/downloads/texas/do"
. * Estimation 1: Texas model of black male prisoners (per capita)
. scuse texas.dta, replace
. ssc install synth.                                     #delimit;
. synth bmprison
. bmprison(1990) bmprison(1992) bmprison(1991) bmprison(1988)
. alcohol(1990) aidscapita(1990) aidscapita(1991)
. income ur poverty black(1990) black(1991) black(1992)
. perc1519(1990)

. ,
. trunit(48) trperiod(1993) unitnames(state)
. mspeperiod(1985(1)1993) resultsperiod(1985(1)2000)
. keep(..//data/synth/synth_bmprate.dta) replace fig;
. mat list e(V_matrix);
. #delimit cr
. graph save Graph ..//Figures/synth_tx.gph, replace
```

Note that on the first line, you will need to change the path directory, but otherwise, it should run because I'm using standard Unix/DOS notation that allows you to back up and redirect to a different subdirectory using the “..//” command. Now in this example, there's a lot of syntax, so let me walk you through it.

First, you need to install the data from my website using `scuse`. Second, I personally prefer to make the delimiter a semicolon because I want to have all syntax for `synth` on the same screen. I'm

more of a visual person, so that helps me. Next the synth syntax. The syntax goes like this: call synth, then call the outcome variable (bmrison), then the variables you want to match on. Notice that you can choose either to match on the entire pre-treatment average, or you can choose particular years. I choose both. Also recall that [Abadie et al. \[2010\]](#) notes the importance of controlling for pre-treatment outcomes to soak up the heterogeneity; I do that here as well. Once you've listed your covariates, you use a comma to move to Stata options. You first have to specify the treatment unit. The FIPS code for Texas is a 48, hence the 48. You then specify the treatment period, which is 1993. You list the period of time which will be used to minimize the mean squared prediction error, as well as what years to display. Stata will produce both a figure as well as a dataset with information used to create the figure. It will also list the V matrix. Finally, I change the delimiter back to carriage return, and save the figure in the /Figures subdirectory. Let's look at what these lines made ([Figure 107](#)). This is the kind of outcome that you ideally

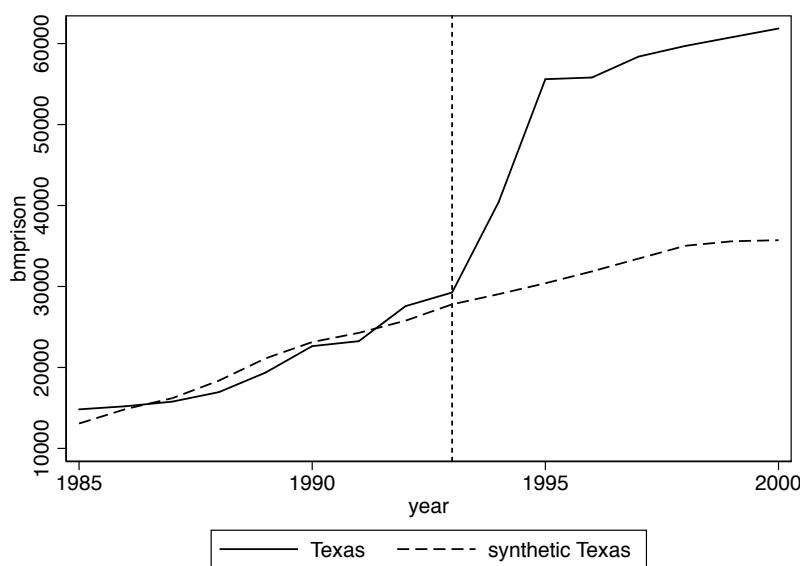


Figure 107: African-American male incarceration

want to say – specifically, a very similar pre-treatment trend in the synthetic Texas group compared to the actual Texas group, and a divergence in the post-treatment period. We will now plot the gap

between these two lines using the following commands:

```
* Plot the gap in predicted error
.use ../data/synth/synth_bmprate.dta, clear
.keep _Y_treated _Y_synthetic _time
.drop if _time==.
.rename _time year
.rename _Y_treated treat
.rename _Y_synthetic counterfactual
.gen gap48=treat-counterfactual
.sort year
.#delimit ;
.twoway (line gap48 year,lp(solid)lw(vthin)lcolor(black)), yline(0, lpattern(shortdash) lcolor(black))
.xline(1993, lpattern(shortdash) lcolor(black)) xtitle("",si(medsmall)) xlabel(#10)
.ytitle("Gap in black male prisoner prediction error", size(medsmall)) legend(off);
.#delimit cr
.save ../data/synth/synth_bmprate_48.dta, replace
```

The figure that this makes is basically nothing more than the gap between the actual Texas and the synthetic Texas from Figure 107.

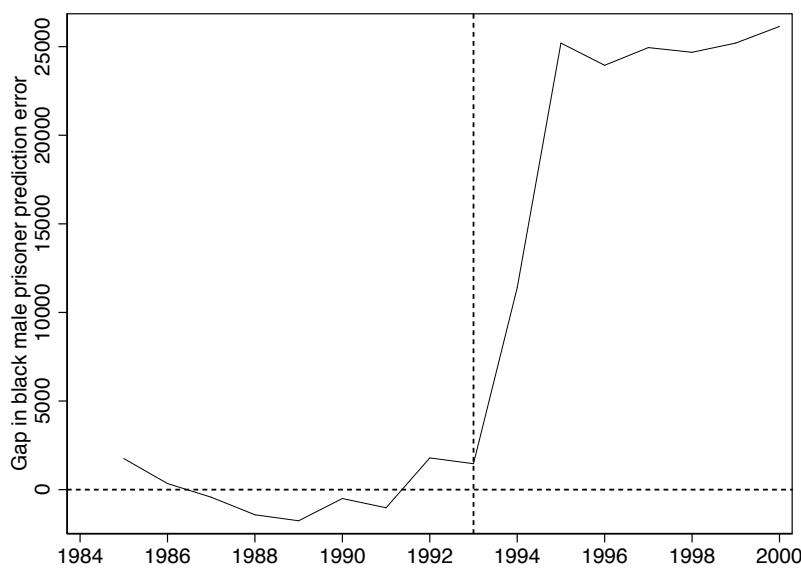


Figure 108: Gap between actual Texas and synthetic Texas

And finally, we will show the weights used to construct the synthetic Texas.

State name	Weight
California	0.408
Florida	0.109
Illinois	0.36
Louisiana	0.122

Table 36: Synthetic control weights

Now that we have our estimates of the causal effect, we move into the calculation of the exact p -value which will be based on assigning the treatment to every state and re-estimating our model. Texas will always be thrown back into the donor pool each time.

```
.* Inference 1 placebo test
.#delimit;
.set more off;
.use ../data/texas.dta, replace;
.local statelist 1 2 4 5 6 8 9 10 11 12 13 15 16 17 18 20 21 22 23 24 25 26 27 28 29 30 31 32
    33 34 35 36 37 38 39 40 41 42 45 46 47 48 49 51 53 55;
.foreach i of local statelist {};
.synth bmprison
. bmprison(1990) bmprison(1992) bmprison(1991) bmprison(1988)
. alcohol(1990) aidscapita(1990) aidscapita(1991)
. income ur poverty black(1990) black(1991) black(1992)
. perc1519(1990)

.,
. trunit('i') trperiod(1993) unitnames(state)
. mspeperiod(1985(1)1993) resultsperiod(1985(1)2000)
. keep(../data/synth/synth_bmrate_`i'.dta) replace;
. matrix state`i' = e(RMSPE); /* check the V matrix*/
. };
.foreach i of local statelist {};
.matrix rownames state`i'='i';
.matlist state`i', names(rows);
.};
.#delimit cr
```

This is a loop in which it will cycle through every state and estimate the model. It will then save data associated with each model into the ..//data/synth/synth_bmrate_`i'.dta data file where `i' is one of the state FIPS code listed after local statelist. Now that we have each

of these files, we can calculate the post-to-pre RMSPE.

```
.local statelist 1 2 4 5 6 8 9 10 11 12 13 15 16 17 18 20 21 22 23 24 25 26 27 28 29 30 31
32 33 34 35 36 37 38 39 40 41 42 45 46 47 48 49 51 53 55
. foreach i of local statelist {
. use ../data/synth/synth_bmprate`i' ,clear
. keep _Y_treated _Y_synthetic _time
. drop if _time==.
. rename _time year
. rename _Y_treated treat`i'
. rename _Y_synthetic counterfact`i'
. gen gap`i'=treat`i'-counterfact`i'
. sort year
. save ../data/synth/synth_gap_bmprate`i', replace
.use ../data/synth/synth_gap_bmprate48.dta, clear
.sort year
.save ../data/synth/placebo_bmprate48.dta, replace
.foreach i of local statelist {
. merge year using ../data/synth/synth_gap_bmprate`i'
. drop _merge
. sort year
. save ../data/synth/placebo_bmprate.dta, replace
```

Notice that this is going to first create the gap between the treatment state and the counterfactual state before merging each of them into

single data file.

```
** Inference 2: Estimate the pre- and post-RMSPE and calculate the ratio of the
.* post-pre RMSPE
.set more off
.local statelist 1 2 4 5 6 8 9 10 11 12 13 15 16 17 18 20 21 22 23 24 25 26 27 28 29 30 31
32 33 34 35 36 37 38 39 40 41 42 45 46 47 48 49 51 53 55
.foreach i of local statelist {
    .use ../data/synth/synth_gap_bmprate'i', clear
    .gen gap3=gap'i'*gap'i'
    .egen postmean=mean(gap3) if year>1993
    .egen premean=mean(gap3) if year<=1993
    .gen rmspe=sqrt(premean) if year<=1993
    .replace rmspe=sqrt(postmean) if year>1993
    .gen ratio=rmspe/rmspe[_n-1] if year==1994
    .gen rmspe_post=sqrt(postmean) if year>1993
    .gen rmspe_pre=rmspe[_n-1] if year==1994
    .mkmatrix rmspe_pre rmspe_post ratio if year==1994, matrix (state'i')
}
```

In this part, we are calculating the post-RMSPE, the pre-RMSPE and the ratio of the two. Once we have this information, we can compute a histogram. The following commands do that.

```

.* show post/pre-expansion RMSPE ratio for all states, generate histogram
.foreach i of local statelist {
    . matrix rownames state'i'='i'
    . matlist state'i', names(rows)
}
.#delimit ;
. mat state=state1\state2\state4\state5\state6\state8\state9\state10\state11\state12\state13\state15
. \state16\state17\state18\state20\state21\state22\state23\state24\state25\state26
. \state27\state28\state29\state30\state31\state32\state33\state34\state35\state36\
. state37\state38\state39\state40\state41\state42\state45\state46\state47\state48;
. \state49\state51\state53\state55;
.#delimit cr
ssc install mat2txt
. mat2txt, matrix(state) saving(../inference/rmspe_bmrate.txt) replace
. insheet using ..//inference/rmspe_bmrate.txt, clear
. ren v1 state
. drop v5
. gsort -ratio
. gen rank=_n
. gen p=rank/46
. export excel using ..//inference/rmspe_bmrate, firstrow(variables) replace
. import excel ..//inference/rmspe_bmrate.xls, sheet("Sheet1") firstrow clear
. histogram ratio, bin(20) frequency fcolor(gs13) lcolor(black) ylabel(0(2)6) xtitle(Post/pre RMSPE ratio)
.* Show the post/pre RMSPE ratio for all states, generate the histogram.
.list rank p if state==48

```

All the looping will take a few moments to run, but once it is done, it will produce a histogram of the distribution of ratios of post-RMSPE to pre-RMSPE. As you can see from the *p*-value, Texas has the second highest ratio out of 46 state units, giving it a *p*-value of 0.04. We can see that in Figure 110. Notice that in addition to the figure, this created an excel spreadsheet containing information on the pre-RMSPE, the post-RMSPE, the ratio, and the rank. We will want to use that again when we limit our display next to states whose pre-RMSPE are similar to that of Texas.

All the looping will take a few moments to run, but once it is done, it will produce a histogram of the distribution of ratios of post-RMSPE to pre-RMSPE. As you can see from the *p*-value, Texas

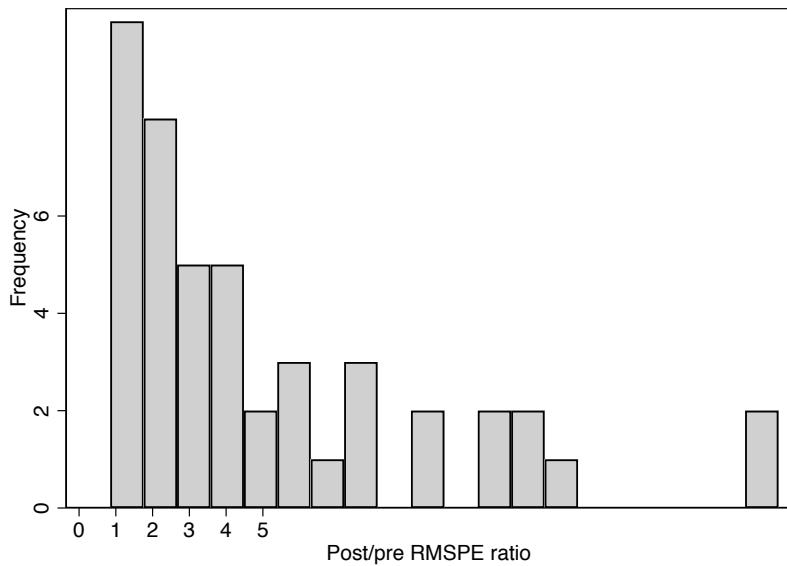


Figure 109: Histogram of the distribution of ratios of post-RMSPE to pre-RMSPE. Texas is one of the ones in the far right tail.

has the second highest ratio out of 46 state units, giving it a p -value of 0.04. We can see that in Figure 110. Notice that in addition to the figure, this created an excel spreadsheet containing information on the pre-RMSPE, the post-RMSPE, the ratio, and the rank. We will want to use that again when we limit our display next to states whose pre-RMSPE are similar to that of Texas.

Now we want to create the characteristic placebo graph where all the state placebos are laid on top of Texas. To do that we use the

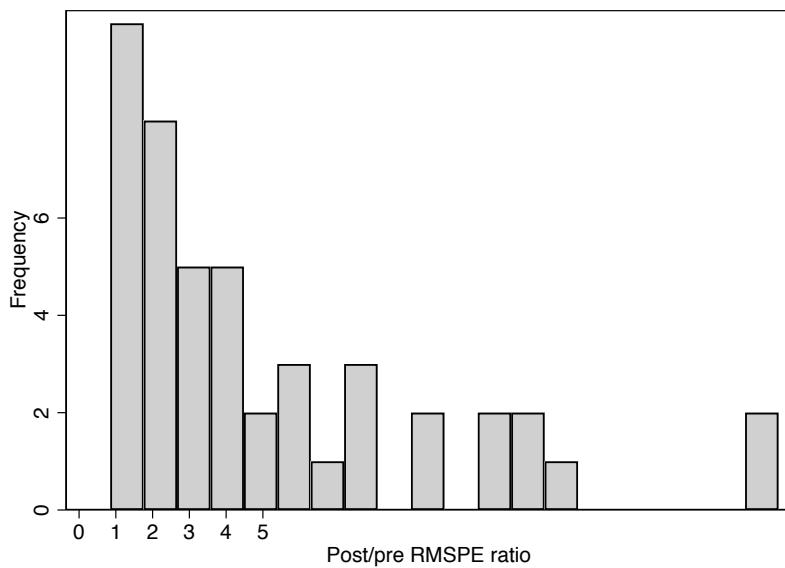


Figure 110: Histogram of the distribution of ratios of post-RMSPE to pre-RMSPE. Texas is one of the ones in the far right tail.

following syntax:

```
.* Inference 3: all the placebos on the same picture
.use ../data/synth/placebo_bmratae.dta, replace
.* Picture of the full sample, including outlier RSMPE
.#delimit;
.twoway
.(line gap1 year ,lp(solid)lw(vthin))
.(line gap2 year ,lp(solid)lw(vthin))
.(line gap4 year ,lp(solid)lw(vthin))
.(line gap5 year ,lp(solid)lw(vthin))
.(line gap6 year ,lp(solid)lw(vthin))
.(line gap8 year ,lp(solid)lw(vthin))
.(line gap9 year ,lp(solid)lw(vthin))
.(line gap10 year ,lp(solid)lw(vthin))
.(line gap11 year ,lp(solid)lw(vthin))
.(line gap12 year ,lp(solid)lw(vthin))
.(line gap13 year ,lp(solid)lw(vthin))
.(line gap15 year ,lp(solid)lw(vthin))
.(line gap16 year ,lp(solid)lw(vthin))
.(line gap17 year ,lp(solid)lw(vthin))
.(line gap18 year ,lp(solid)lw(vthin))
.(line gap20 year ,lp(solid)lw(vthin))
.(line gap21 year ,lp(solid)lw(vthin))
.(line gap22 year ,lp(solid)lw(vthin))
.(line gap23 year ,lp(solid)lw(vthin))
.(line gap24 year ,lp(solid)lw(vthin))
```

```

.(line gap25 year ,lp(solid)lw(vthin))
.(line gap26 year ,lp(solid)lw(vthin))
.(line gap27 year ,lp(solid)lw(vthin))
.(line gap28 year ,lp(solid)lw(vthin))
.(line gap29 year ,lp(solid)lw(vthin))
.(line gap30 year ,lp(solid)lw(vthin))
.(line gap31 year ,lp(solid)lw(vthin))
.(line gap32 year ,lp(solid)lw(vthin))
.(line gap33 year ,lp(solid)lw(vthin))
.(line gap34 year ,lp(solid)lw(vthin))
.(line gap35 year ,lp(solid)lw(vthin))
.(line gap36 year ,lp(solid)lw(vthin))
.(line gap37 year ,lp(solid)lw(vthin))
.(line gap38 year ,lp(solid)lw(vthin))
.(line gap39 year ,lp(solid)lw(vthin))
.(line gap40 year ,lp(solid)lw(vthin))
.(line gap41 year ,lp(solid)lw(vthin))
.(line gap42 year ,lp(solid)lw(vthin))
.(line gap45 year ,lp(solid)lw(vthin))
.(line gap46 year ,lp(solid)lw(vthin))
.(line gap47 year ,lp(solid)lw(vthin))
.(line gap49 year ,lp(solid)lw(vthin))
.(line gap51 year ,lp(solid)lw(vthin))
.(line gap53 year ,lp(solid)lw(vthin))
.(line gap55 year ,lp(solid)lw(vthin))
.(line gap48 year ,lp(solid)lw(thick)\color(black)), /*treatment unit, Texas*/
.yline(0, lpattern(shortdash) \color(black)) xline(1993, lpattern(shortdash) \color(black))
.xtitle("",si(small)) xlabel(#10) ytitle("Gap in black male prisoners prediction error", size(small))
. legend(off);
.#delimit cr

```

Here we will only display the main picture with the placebos, though one could show several cuts of the data in which you drop states whose pre-treatment fit compared to Texas is rather poor. Now that you have seen how to use this do file to estimate a synthetic control model, you are ready to play around with the data yourself. All of this analysis so far has used black male (total counts) incarceration as the dependent variable, but perhaps the results

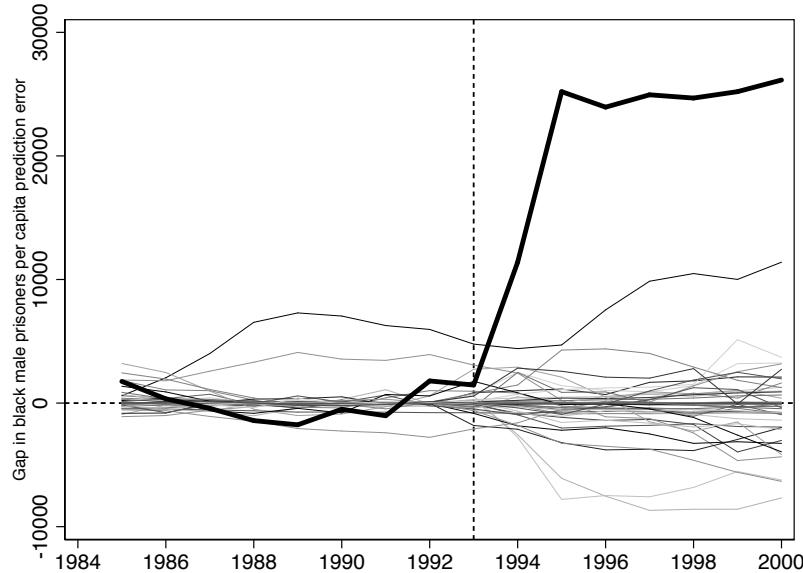


Figure 111: Placebo distribution. Texas is the black line.

would be different if we used black male incarceration rates. That information is contained in the dataset. I would like for you to do your own analysis using the black male incarceration rate variable as the dependent variable. You will need to find a new model to fit this pattern, as it's unlikely that the one we used for levels will do as good a job describing rates as it did levels. In addition, you should implement the placebo-date falsification exercise that we mentioned from [Abadie et al. \[2015\]](#). Choose an 1989 as your treatment date and 1992 as the end of the sample and check whether the same model shows the same treatment effect as you found when you used the correct year, 1993, as the treatment date. I encourage you to use these data and this file to learn the ins and outs of the procedure itself, as well as to think more deeply about what synthetic control is doing and how to best use it in research.

Conclusion In conclusion, we have seen how to estimate synthetic control models in Stata. This model is currently an active area of research (e.g., [Powell \[2017\]](#)), but this is a good foundation for understanding the model. I hope that you find this useful.

Conclusion

Causal inference is a fun area. It's fun because the Rubin causal model is such a philosophically stimulating and intuitive way to think about causal effects, and Pearl's directed acylic graphical models are so helpful for moving between a theoretical model and/or an understanding of some phenomena, and an identification strategy to identify the causal effect you care about. From those DAGs, you will learn whether it's even possible to design such an identification strategy with the dataset you have, and while that can be disappointing, it is nonetheless a disciplined and truthful approach to identification. These DAGs are, in my experience, empowering and extremely useful for the design phase of a project.

The methods I've outlined are merely some of the most common research designs currently employed in applied microeconomics. They are not all methods, and each method is not exhaustively plumbed either. Version 1.0 omits a lot of things, like I said in the opening chapter, such as machine learning, imperfect controls, matrix completion, and structural estimation. I do not omit these because they are unimportant; I omit them because I am still learning them myself!

Version 2.0 will differ from version 1.0 primarily in that it will add in some of these additional estimators and strategies. Version 2.0 will also contain more Stata exercises, and most likely I will produce a set of do files for you that will exactly reproduce the examples I go through in the book. It may be helpful for you to have handy a file, as well as see the programming on the page. I also would like to have more simulation, as I find that simulations are a great way to communicate the identifying assumptions for some estimator, as well as explain basic ideas like the variance in some estimator.

I hope you find this book valuable. Please check out the many papers I've cited, as well as the textbooks I listed at the beginning, as they are all excellent, and you will learn more from them than you have learned from my introductory book. Good luck in your research.

Bibliography

Alberto Abadie and Javier Gardeazabal. The economic costs of conflict: A case study of the basque country. *American Economic Review*, 93(1):113–132, March 2003.

Alberto Abadie and Guido Imbens. Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267, 2006.

Alberto Abadie and Guido Imbens. Bias-corrected matching estimators for average treatment effects. *Journal of Business and Economic Statistics*, 29:1–11, 2011.

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of california’s tobacco control program. *Journal of the American Statistical Association*, 105(490):493–505, June 2010.

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Comparative politics and the synthetic control method. *American Journal of Political Science*, 59(2):495–510, October 2015. Unpublished Manuscript.

Eric Allen, Patricia Dechow, Devin Pope, and George Wu. Reference-dependent preferences: Evidence from marathon runners. Unpublished Manuscript, 2013.

Douglas Almond, Joseph J. Doyle, Amanda Kowalski, and Heidi Williams. Estimating returns to medical care: Evidence from at-risk newborns. *The Quarterly Journal of Economics*, 125(2):591–634, 20010.

Joshua D. Angrist. Lifetime earnings and the vietnam era draft lottery: Evidence from social security administrative records. *American Economic Review*, 80(3):313–336, June 1990.

Joshua D. Angrist and Alan B. Krueger. Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics*, 106(4):979–1014, November 1991.

Joshua D. Angrist and Alan B. Krueger. Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives*, 15(4):69–85, 2001.

Joshua D. Angrist and Victor Lavy. Using maimonides' rule to estimate the effect of class size on scholastic achievement. *Quarterly Journal of Economics*, 114(2):533–575, 1999.

Joshua D. Angrist and Jorn-Steffen Pischke. *Mostly Harmless Econometrics*. Princeton University Press, 1st edition, 2009.

Joshua D. Angrist, Guido W. Imbens, and Donald B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 87:328–336, 1996.

Susan Athey and Guide W. Imbens. The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, 31(2):3–32, Spring 2017.

Susan Athey, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi. Matrix completion methods for causal panel data models. Unpublished Manuscript, October 2017.

M. Christopher Auld and Paul Grootendorst. An empirical analysis of milk addiction. *Journal of Health Economics*, 23(6):1117–1133, November 2004.

David H. Autor. Outsourcing at will: The contribution of unjust dismissal doctrine to the growth of employment outsourcing. *Journal of Labor Economics*, 21(1):1–42, 2003.

Katherine Baicker, Sarah L. Taubman, Heidi L. Allen, Mira Bernstein, Jonathan Gruber, Joseph Newhouse, Eric Schneider, Bill Wright, Alam Zaslavsky, and Amy Finkelstein. The oregon experiment – effects of medicaid on clinical outcomes. *New England Journal of Medicine*, 368:1713–1722, May 2013.

Burt S. Barnow, Glen G. Cain, and Arthur Goldberger. Selection on observables. *Evaluation Studies Review Annual*, 5:43–59, 1981.

Gary Becker. Crime and punishment: An economic approach. *The Journal of Political Economy*, 76:169–217, 1968.

Gary Becker. *Human Capital: A Theoretical and Empirical Analysis with Special Reference to Education*. University of Chicago Press, 3rd edition, 1994.

Gary S. Becker. The economic way of looking at life. University of Chicago Coase-Sandor Working Paper Series in Law and Economics, 1993.

Gary S. Becker and Kevin M. Murphy. A theory of rational addiction. *Journal of Political Economy*, 96(4), August 1988.

Gary S. Becker, Michael Grossman, and Kevin M. Murphy. The market for illegal gods: The case of drugs. *Journal of Political Economy*, 114(1):38–60, 2006.

Marianne Bertrand, Esther Duflo, and Sendhil Mullainathan. How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics*, 119(1):249–275, February 2004.

Sandra E. Black. Do better schools matter? parental valuation of elementary education. *Quarterly Journal of Economics*, 114(2):577–599, 1999.

John Bound, David A. Jaeger, and Regina M. Baker. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association*, 90(430), 1995.

John M. Brooks and Robert L. Ohsfeldt. Squeezing the balloon: Propensity scores and unmeasured covariate balance. *Health Services Research*, 48(4):1487–1507, August 2013.

Sebastian Calonico, Matis D. Cattaneo, and Rocio Titiunik. Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, 82(6):2295–2326, November 2014.

David Card. The impact of the mariel boatlift on the miami labor market. *Industrial and Labor Relations Review*, 43(2):245–257, January 1990.

David Card. *Aspects of Labour Economics: Essays in Honour of John Vandenkamp*, chapter Using Geographic Variation in College Proximity to Estimate the Return to Schooling. University of Toronto Press, 1995.

David Card and Alan Krueger. Minimum wages and employment: A case study of the fast-food industry in new jersey and pennsylvania. *American Economic Review*, 84:772–793, 1994.

David Card and Giovanni Peri. Immigration economics: A review. Unpublished Manuscript, April 2016.

David Card, Carlos Dobkin, and Nicole Maestas. The impact of nearly universal insurance coverage on health care utilization: Evidence from medicare. *American Economic Review*, 98(5):2242–2258, December 2008.

David Card, Carlos Dobkin, and Nicole Maestas. Does medicare save lives? *The Quarterly Journal of Economics*, 124(2):597–636, 2009.

David Card, David S. Lee, Zhuan Pei, and Andrea Weber. Inference on causal effects in a generalized regression kink design. *Econometrica*, 84(6):2453–2483, November 2015.

Christopher Carpenter and Carlos Dobkin. The effect of alcohol consumption on mortality: Regression discontinuity evidence from the minimum drinking age. *American Economic Journal: Applied Economics*, 1(1):164–182, January 2009.

Scott E. Carrell, Mark Hoekstra, and James E. West. Does drinking impair college performance? evidence from a regression discontinuity approach. *Journal of Public Economics*, 95:54–62, 2011.

Eduardo Cavallo, Sebastian Galiani, Ilan Noy, and Juan Pantano. Catastrophic natural disasters and economic growth. *Review of Economics and Statistics*, 95(5):1549–1561, 2013.

Kerwin Charles and Ming Ching Luoh. Male incarceration, the marriage market and female outcomes. Unpublished Manuscript, 2006.

Cheng Cheng and Mark Hoekstra. Does strengthening self-defense law deter crime or escalate violence? evidence from expansions to castle doctrine. *Journal of Human Resources*, 48(3):821–854, 2013.

W. G. Cochran. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, 24(2):295–313, 1968.

Ethan Cohen-Cole and Jason Fletcher. Detecting implausible social network effects in acne, height, and headaches: Longitudinal analysis. *British Medical Journal*, 337(a2533), 2008.

Dalton Conley and Jason Fletcher. *The Genome Factor: What the Social Genomics Revolution Reveals about Ourselves, Our History, and the Future*. Princeton University Press, 2017.

Thomas D. Cook. “waiting for life to arrive”: A history of the regression-discontinuity design in psychology, statistics and economics. *Journal of Econometrics*, 142:636–654, 2008.

Christopher Cornwell and Scott Cunningham. Mass incarceration’s effect on risky sex. Unpublished Manuscript, 2016.

Christopher Cornwell and Peter Rupert. Unobservable individual effects, marriage and the earnings of young men. *Economic Inquiry*, 35(2):1–8, April 1997.

Christopher Cornwell and William N. Trumbull. Estimating the economic model of crime with panel data. *Review of Economics and Statistics*, 76(2):360–366, 1994.

Michael Craig. *The Professor, the Banker and the Suicide King: Inside the Richest Poker Game of All Time*. Grand Central Publishing, 2006.

Richard K. Crump, V. Joseph Hotz, Guido W. Imbens, and Oscar A. Mitnik. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–1999, 2009.

Scott Cunningham and Christopher Cornwell. The long-run effect of abortion on sexually transmitted infections. *American Law and Economics Review*, 15(1):381–407, Spring 2013.

Scott Cunningham and Keith Finlay. Parental substance abuse and foster care: Evidence from two methamphetamine supply shocks? *Economic Inquiry*, 51(1):764–782, 2012.

Scott Cunningham and Todd D. Kendall. Prostitution 2.0: The changing face of sex work. *Journal of Urban Economics*, 69:273–287, 2011.

Scott Cunningham and Todd D. Kendall. Prostitution labor supply and education. *Review of Economics of the Household*, Forthcoming, 2016.

Stacy Berg Dale and Alan B. Krueger. Estimating the payoff to attending a more selective college: An application of selection on observables and unobservables. *Quarterly Journal of Economics*, 117(4):1491–1527, November 2002.

Angus Deaton and Nancy Cartwright. Understanding and misunderstanding randomized controlled trials. *Social Science and Medicine*, forthcoming, 2018.

Rajeev H. Dehejia and Sadek Wahba. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94(448):1053–1062, December 1999.

Rajeev H. Dehejia and Sadek Wahba. Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics*, 84(1):151–161, February 2002.

Carlos Dobkin and Nancy Nicosia. The war on drugs: Methamphetamine, public health and crime. *American Economic Review*, 99(1):324–349, 2009.

John J. Donohue and Steven D. Levitt. The impact of legalized abortion on crime. *The Quarterly Journal of Economics*, 116(2):379–420, May 2001.

Nikolay Doudchenko and Guido Imbens. Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. NBER Working Papers 22791, 2016.

Mirko Draca, Stephen Machin, and Robert Witt. Panic on the streets of london: Police, crime, and the july 2005 terror attacks. *American Economic Review*, 101(5):2157–81, August 2011.

Arindrajit Dube, T. William Lester, and Michael Reich. Minimum wage effects across state borders: Estimates using contiguous counties. *Review of Economics and Statistics*, 92(4):945–964, November 2010.

William N. Evans and Emily G. Owens. Cops and crime. *Journal of Public Economics*, 91(1-2):181–201, February 2007.

R. A. Fisher. *The Design of Experiments*. Edinburgh: Oliver and Boyd, 1935.

Roland A. Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburg, 1925.

Christopher L. Foote and Christopher F. Goetz. The impact of legalized abortion on crime: Comment. *Quarterly Journal of Economics*, 123(1):407–423, February 2008.

David A. Freedman. Statistical models and shoe leather. *Sociological Methodology*, 21:291–313, 1991.

Ragnar Frisch and Frederick V. Waugh. Partial time regressions as compared with individuals trends. *Econometrica*, 1(4):387–401, 1933.

Carl Friedrich Gauss. *Theoria Motus Corporum Coelestium*. Perthes et Besser, Hamburg, 1809.

Andrew Gelman and Guido Imbens. Why higher-order polynomials should not be used in regression discontinuity designs. *Journal of Business and Economic Statistics*, <https://doi.org/10.1080/07350015.2017.1366909>, 2017.

Andrew Gelman and Guido W. Imbens. Why high-order polynomials should not be used in regression discontinuity design. Unpublished Manuscript, September 2016.

Paul Gertler, Manisha Shah, and Stefano M. Bertozzi. Risky business: The market for unprotected commercial sex. *Journal of Political Economy*, 113(3):518–550, 2005.

- A. S. Goldberger. Selection bias in evaluating treatment effects: some formal illustrations. Madison, WI unpublished Manuscript, 1972.
- Kathryn Graddy. The fulton fish market. *Journal of Economic Perspectives*, 20(2):207–220, Spring 2006.
- Jonathan Gruber. The incidence of mandated maternity benefits. *American Economic Review*, 84(3):622–641, June 1994.
- Jonathan Gruber, Phillip B. Levine, and Douglas Staiger. Abortion legalization and child living circumstances: Who is the “marginal child”? *The Quarterly Journal of Economics*, 114(1):263–291, February 1999.
- Trygve Haavelmo. The statistical implications of a system of simultaneous equations. *Econometrica*, 11(1):1–12, 1943.
- Jinyong Hahn, Petra Todd, and Wilbert van der Klaauw. Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, 69(1):201–209, January 2001.
- Daniel S. Hamermesh and Jeff E. Biddle. Beauty and the labor market. *American Economic Review*, 84(5):1174–1194, 1994.
- Ben Hansen. Punishment and deterrence: Evidence from drunk driving. *American Economic Review*, 105(4):1581–1617, 2015.
- James Heckman and Rodrigo Pinto. Causal analysis after haavelmo. *Econometric Theory*, 31(1):115–151, February 2015.
- James J. Heckman and Edward J. Vytlacil. *Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation*, volume 6B, chapter 70, pages 4779 – 4874. Elsevier, 2007.
- Wayne H. Holtzman. The unbiased estimate of the population variance and standard deviation. *The American Journal of Psychology*, 63(4):615–617, 1950.
- Seung-Hyun Hong. Measuring the effect of napster on recorded music sales: Difference-in-differences estimates under compositional changes. *Journal of Applied Econometrics*, 28(2):297–324, March 2013.
- Robert Hooke. *How to Tell the Liars from the Statisticians*. CRC Press, 1983.
- David Hume. *An Enquiry Concerning Human Understanding: with Hume's Abstract of A Treatise of Human Nature and A Letter from a Gentleman to His Friend in Edinburgh*. Hackett Publishing Company, 2nd edition, 1993.

Stefano M. Iacus, Gary King, and Giuseppe Porro. Causal inference without balance checking: Coarsened exact matching. *Political Analysis*, 20(1):1–24, 2012.

Kosuke Imai and In Song Kim. When should we use fixed effects regression models for causal inference with longitudinal data. Unpublished Manuscript, December 2017.

Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social and Biomedical Sciences: An Introduction*. Cambridge University Press, 1st edition, 2015.

Guido W. Imbens and Joshua D. Angrist. Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475, 1994.

Guido Imbens. Better late than nothing: Some comments on deaton (2009) and heckman and urzua (2009). Unpublished Manuscript, April 2009.

Guido Imbens and Karthik Kalyanaraman. Optimal bandwidth choice for the regression discontinuity estimator. *The Review of Economic Studies*, 79(3):933–959, July 2011.

Guido W. Imbens and Thomas Lemieux. Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142:615–635, 2008.

Brian A. Jacob and Lars Lefgen. Remedial education and student achievement: A regression-discontinuity analysis. *The Review of Economics and Statistics*, 86(1):226–244, February 2004.

Ted Joyce. Did legalized abortion lower crime? *The Journal of Human Resources*, 39(1):1–28, Winter 2004.

Ted Joyce. A simple test of abortion and crime. *Review of Economics and Statistics*, 91(1):112–123, 2009.

Chinhui Juhn, Kevin M. Murphy, and Brooks Pierce. Wage inequality and the rise in returns to skill. *Journal of Political Economy*, 101(3):410–442, June 1993.

Gary King and Langche Zeng. The dangers of extreme counterfactuals. *Political Analysis*, 14(2):131–159, 2006.

Alan Krueger. Experimental estimates of education production functions. *Quarterly Journal of Economics*, 114(2):497–532, May 1999.

Robert Lalonde. Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review*, 76(4):604–620, 1986.

David S. Lee. Randomized experiments from non-random selection in u.s. house elections. *Journal of Econometrics*, 142:675–697, 2008.

David S. Lee and Thomas Lemieux. Regresion discontinuity designs in economics. *Journal of Economic Literature*, 48:281–355, June 2010.

David S. Lee, Enrico Moretti, and Matthew J. Butler. Do voters affect or elect policies: Evidence from the u.s. house. *Quarterly Journal of Economics*, 119(3):807–859, August 2004.

Phillip B. Levine. *Sex and Consequences: Abortion, Public Policy, and the Economics of Fertility*. Princeton University Press, 1st edition, 2004.

Phillip B. Levine, Douglas Staiger, Thomas J. Kane, and David J. Zimmerman. Roe v. wade and american fertility. *American Journal of Public Health*, 89(2):199–203, February 1999.

Steven D. Levitt. Understanding why crime fell in the 1990s: Four factors that explain the decline and six that do not. *Journal of Economic Perspectives*, 18(1):163–190, Winter 2004.

David Lewis. Causation. *The Journal of Philosophy*, 70(17):556–567, October 1973.

John R. Lott and David B. Mustard. Crime, deterrence and the right-to-carry concealed handguns. *Journal of Legal Studies*, 26:1–68, 1997.

Michael C. Lovell. Seasonal adjustment of economic time series and multiple regression analysis. *Journal of the American Statistical Association*, 58(304):991–1010, 1963.

Michael C. Lovell. A simple proof of the fwl theorem. *Journal of Economic Education*, 39(1):88–91, 2008.

Ross L. Matsueda. *Handbook of Structural Equation Modeling*, chapter “Key Advances in the History of Structural Equation Modeling”. Guilford Press, 2012.

Justin McCrary. Manipulation of the running variable in the regression discontinuity design: A design test. *Journal of Econometrics*, 142:698–714, 2008.

John Stuart Mill. *A System of Logic, Ratiocinative and Inductive*. FQ Books, July 2010.

Mary S. Morgan. *The History of Econometric Ideas*. Cambridge University Press, 1991.

Stephen L. Morgan and Christopher Winship. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge University Press, 2nd edition, 2014.

Martin Needleman and Carolyn Needleman. Marx and the problem of causation. *Science and Society*, 33(3):322–339, Summer - Fall 1969.

David Neumark, J.M. Ian Salas, and William Wascher. Revisting the minimum wage-employment debate: Throwing out the baby with the bathwater? *Industrial and Labor Relations Review*, 67(2.5):608–648, 2014.

Joseph P. Newhouse. *Free for All? Lessons from the RAND Health Experiment*. Harvard University Press, 1993.

Judea Pearl. *Causality*. Cambridge University Press, 2nd edition, 2009.

Charles Sanders Peirce and Joseph Jastrow. On small differences in sensation. *Memoirs of the National Academy of Sciences*, 3:73–83, 1885.

Giovanni Peri and Vasil Yashenov. The labor market effects of a refugee wave: Synthetic control method meets the mariel boatlift. *Journal of Human Resources*, doi: 10.3388/jhr.54.2.0217.8561R1, 2018 2018.

Robert Perkinson. *Texas Tough: The Rise of America's Prison Empire*. Picador, first edition, 2010.

David Powell. Imperfect synthetic controls: Did the massachusetts health care reform save lives? Unpublished Manuscript, October 2017.

Ranier Maria Rilke. *Letters to a Young Poet*. Merchant Books, 2012.

Sherwin Rosen. *Handbook of Labor Economics*, volume 1, chapter The Theory of Equalizing Differences. Amsterdam: North-Holland, 1986.

Paul R. Rosenbaum. Two simple models for observational studies. *Design of Observational Studies*, pages 65–94, 2010.

Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, April 1983.

Donald Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

Donald B. Rubin. Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics*, 2:1–26, 1977.

- Donald B. Rubin. Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics*, 31:161–170, 2004.
- Donald B. Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, March 2005.
- John Rust. Optimal replacement of gmc bus engines: An empirical model of harold zurcher. *Econometrica*, 55(5):999–1033, 1987.
- Adam Smith. *The Wealth of Nations*. Bantam Classics, 2003.
- Jerzy Splawa-Neyman. On the application of probability theory to agricultural experiments. essay on principles. *Annals of Agricultural Sciences*, pages 1–51, 1923.
- Douglas Staiger and James H. Stock. Instrumental variables regression with weak instruments. *Econometrica*, 65(3):557–586, 1997.
- James H. Stock and Francesco Trebbi. Who invented instrumental variable regression? *The Journal of Economic Perspectives*, 17(3):177–194, Summer 2003.
- James H. Stock and Motohiro Yogo. Testing for weak instruments in linear iv regression. In Donald W. K. Andrews and James H. Stock, editors, *Identification and Inference for Econometrics Models: Essays in Honor of Thomas Rothenberg*. Cambridge University Press, 2005.
- Jack Stuster and Marcelline Burns. Validation of the standardized field sobriety test battery at bacs below 0.10 percent. Technical report, US Department of Transportation, National Highway Traffic Safety Administration, August 1998.
- Donald Thistlewaite and Donald Campbell. Regression-discontinuity analysis: an alternative to the ex-post facto experiment. *Journal of Educational Psychology*, 51:309–317, 1960.
- Wilbert van der Klaauw. Estimating the effect of financial aid offers on college enrollment: A regression-discontinuity approach. *International Economic Review*, 43(4):1249–1287, November 2002.
- Jeffrey Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. MIT Press, 2nd edition, 2010.
- Jeffrey Wooldridge. *Introductory Econometrics: A Modern Approach*. South-Western College Pub, 6th edition, 2015.
- Phillip G. Wright. *The Tariff on Animal and Vegetable Oils*. The Macmillan Company, 1928.

G. Udny Yule. An investigation into the causes of changes in pauperism in england, chiefly during the last two interensal decades.
Journal of Royal Statistical Society, 62:249–295, 1899.