Causal Inference

Scott Cunningham

Causal Inference
The Mixtape

Causal Inference

The Mixtape

Scott Cunningham

Yale
UNIVERSITY PRESS
NEW HAVEN & LONDON

Set in Roboto type by Newgen.

Title-page illustration: iStock.com/2p2play.

Printed in the United States of America.

Library of Congress Control Number: 2020939011

ISBN 978-0-300-25168-5 (pbk. : alk. paper).

A catalogue record for this book is available from the British Library.

This paper meets the requirements of ANSI/NISO Z39.48-1992 (Permanence of Paper).

10 9 8 7 6 5 4 3 2 1

To my son, Miles, one of my favorite people.

I love you. You've tagged my head and heart.

Contents

Acknowledgments

Just as it takes a village to raise a child, it takes many people to help me write a book like this. The people to whom I am indebted range from the scholars whose work has inspired me—Alberto Abadie, Josh Angrist, Susan Athey, David Card, Esther Duflo, Guido Imbens, Alan Krueger, Robert LaLonde, Steven Levitt, Alex Tabarrok, John Snow, and many more—to friends, mentors, and colleagues.

I am most indebted first of all to my former advisor, mentor, coauthor, and friend Christopher Cornwell. I probably owe Chris my entire career. He invested in me and taught me econometrics as well as empirical designs more generally when I was a grad student at the University of Georgia. I was brimming with a million ideas and he some-how managed to keep me focused. Always patient, always holding me to high standards, always believing I could achieve them, always trying to help me correct fallacious reasoning and poor knowledge of econometrics. I would also like to thank Alvin Roth, who has encouraged me over the last decade in my research. That encouragement has buoyed me throughout my career repeatedly. Finally, I'd like to thank Judea Pearl for inviting me to UCLA for a day of discussions around an earlier draft of the Mixtape and helping improve it.

But a book like this is also due to countless conversations with friends over the years, as well as reading carefully their own work and learning from them. People like Mark Hoekstra, Rebecca Thornton, Paul Goldsmith-Pinkham, Mark Anderson, Greg DeAngelo, Manisha Shah, Christine Durrance, Melanie Guldi, Caitlyn Myers, Bernie Black, Keith Finlay, Jason Lindo, Andrew Goodman-Bacon, Pedro Sant'anna, Andrew Baker, Rachael Meager, Nick Papageorge, Grant McDermott, Salvador Lozano, Daniel Millimet, David Jaeger, Berk Ozler, Erin Hengel, Alex Bartik, Megan Stevenson, Nick Huntington-Klein, Peter Hull, as well as many many

more on #EconTwitter, a vibrant community of social scientists on Twitter.

I would also like to thank my two students Hugo Rodrigues and Terry Tsai. Hugo and Terry worked tirelessly to adapt all of my blue collar Stata code into R programs. Without them, I would have been lost. I would also like to thank another student, Brice Green, for early trials of the code to confirm it worked by non-authors. Blagoj Gegov helped create many of the figures in Tikz. I would like to thank Ben Chidmi for adapting a simulation from R into Stata, and Yuki Yanai for allowing me to use his R code for a simulation. Thank you to Zeljko Hrcek for helping make amendments to the formatting of the LaTeX when I was running against deadline. And thank you to my friend Seth Hahne for creating several beautiful illustrations in the book. I would also like to thank Seth Ditchik for believing in this project, my agent Lindsay Edgecombe for her encouragement and work on my behalf, and Yale University Press. And to my other editor, Charlie Clark, who must have personally read this book fifty times and worked so hard to improve it. Thank you, Charlie. And to the musicians who have sung the soundtrack to my life, thanks to Chance, Drake, Dr. Dre, Eminem, Lauryn Hill, House of Pain, Jay-Z, Mos Def, Notorious B.I.G., Pharcyde, Tupac Shakur, Tribe, Kanye West, Young MC, and many others.

Finally, I'd like to thank my close friends, Baylor colleagues, students, and family for tolerating my eccentric enthusiasm for causal inference and economics for years. I have benefited tremendously from many opportunities and resources, and for that and other things I am very grateful.

This book, and the class it was based on, is a distillation of countless journal articles, books, as well as classes I have taken in person and studied from afar. It is also a product of numerous conversations I've had with colleagues, students and teachers for many years. I have attempted to give credit where credit is due. All errors in this book were caused entirely by me, not the people listed above.

Causal Inference

Introduction

My path to economics was not linear. I didn't major in economics, for instance. I didn't even take an economics course in college. I majored in English, for Pete's sake. My ambition was to become a poet. But then I became intrigued with the idea that humans can form plausible beliefs about causal effects even without a randomized experiment. Twenty-five years ago, I wouldn't have had a clue what that sentence even meant, let alone how to do such an experiment. So how did I get here? Maybe you would like to know how I got to the point where I felt I needed to write this book. The TL;DR version is that I followed a windy path from English to causal inference.[1] First, I fell in love with economics. Then I fell in love with empirical research. Then I noticed that a growing interest in causal inference had been happening in me the entire time. But let me tell the longer version.

I majored in English at the University of Tennessee at Knoxville and graduated with a serious ambition to become a professional poet. But, while I had been successful writing poetry in college, I quickly realized that finding 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. Slowly, I had stopped writing poetry altogether.[2]

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 observations. I did this by running focus groups and conducting in-depth interviews, as well as through 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 website when a speech by Gary Becker caught my eye. It was his Nobel Prize acceptance speech on how economics applies 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 an article on crime by Lott and Mustard [1997] that I became truly enamored of economics. I had no idea that there was an empirical component where economists sought to estimate causal effects with quantitative data. A coauthor of that paper was David Mustard, then an associate professor of economics at the University of Georgia, and one of Gary Becker's former students. I decided that I wanted to study with Mustard, and so I applied to the 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 first-year comprehensive exams, I took Mustard's labor economics field class and learned about a variety of topics that would shape my interests for years. These topics included the returns to education, inequality, racial discrimination, crime, and many other fascinating topics in labor. 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 research I cared about. In fact, I decided to make econometrics my main field of study. This led me to work with Christopher Cornwell, an econometrician and labor economist at Georgia. I learned a lot from Chris, both about econometrics and about research itself. He became a mentor, coauthor, and close friend.

Econometrics was difficult. I won't even pretend I was good at it. I took all the econometrics courses offered at the University of Georgia, some more than once. They included classes covering topics like probability and statistics, cross-sections, panel data, time series, and qualitative dependent variables. But while I passed my field exam in econometrics, I struggled to understand econometrics at a deep level. As the saying goes, I could not see the forest for the trees. Something just wasn't clicking.

I noticed something, though, 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 the cohort's future sexual behavior [Cunningham and Cornwell, 2013]. It was a revisiting of Donohue and Levitt [2001]. One of the books I read in preparation for my study was Levine [2004], which in addition to reviewing the theory of and empirical studies on abortion had a little table explaining the difference-in-differences identification strategy. The University of Georgia had a traditional econometrics pedagogy, and most of my field courses were theoretical (e.g., public economics, industrial organization), so I never really had heard the phrase "identification strategy," let alone "causal inference." Levine's simple difference-in-differences table for some reason opened my eyes. I saw how econometric modeling could be used to isolate the causal effects of some treatment, and that led to a change in how I approach empirical problems.

What Is Causal Inference?

My first job out of graduate school was as an assistant professor at Baylor University in Waco, Texas, where I still work and live today. I was restless the second I got there. I could feel that econometrics was indispensable, and yet I was missing something. But what? It was a theory of causality. I had been orbiting that theory ever since seeing that difference-in-differences table in Levine [2004]. But I needed more. So, desperate, I did what I always do when I want to learn something new—I developed a course on causality to force myself to learn all the things I didn't know.

I named the course Causal Inference and Research Design and taught it for the first time to Baylor master's students in 2010. At the time, I couldn't really find an example of the sort of class I was looking for, so I cobbled together a patchwork of ideas from several disciplines and authors, like labor economics, public economics, sociology, political science, epidemiology, and statistics. You name it. My class wasn't a pure econometrics course; rather, it was an applied empirical class that taught a variety of contemporary research designs, such as difference-in-differences, and it was filled with empirical replications and readings, all of which were built on the robust theory of causality found in Donald Rubin's work as well as the work of Judea Pearl. This book and that class are in fact very similar to one another.[3]

So how would I define causal inference? Causal inference is the leveraging of theory and deep knowledge of institutional details to estimate the impact of events and choices on a given outcome of interest. It is not a new field; humans have been obsessing over causality since antiquity. But what is new is the progress we believe we've made in estimating causal effects both inside and outside the laboratory. Some date the beginning of this new, modern causal inference to Fisher [1935], Haavelmo [1943], or Rubin [1974]. Some connect it to the work of early pioneers like John Snow. We should give a lot of credit to numerous highly creative labor economists from the late 1970s to late 1990s whose ambitious research agendas created a revolution in economics that continues to this day. You could even make an argument that we owe it to the Cowles Commission, Philip and Sewall Wright, and the computer scientist Judea Pearl.

But however you date its emergence, causal inference has now matured into a distinct field, and not surprisingly, you're starting to see more and more treatments of it as such. It's sometimes reviewed in a lengthy chapter on "program evaluation" in econometrics textbooks [Wooldridge, 2010], or even 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 those found in Angrist and Krueger [2001] and Imbens and Lemieux [2008]. The market is quietly adding books and articles about identifying causal effects with data all the time.

So why does *Causal Inference: The Mixtape* exist? Well, to put it bluntly, a readable introductory book with programming examples, data, and detailed exposition didn't exist until this one. My book is an effort to fill that hole, because I believe what researchers really need is a guide that takes them from knowing almost nothing about causal inference to a place of competency. Competency in the sense that they are conversant and literate about what designs can and cannot do. Competency in the sense that they can take data, write code and, using theoretical and contextual knowledge, implement a reasonable design in one of their own projects. If this book helps someone do that, then this book will have had value, and that is all I can and should hope for.

But what books out there do I like? Which ones have inspired this book? And why don't I just keep using them? For my classes, I mainly relied on Morgan and Winship [2014], Angrist and Pischke [2009], as well as a library of theoretical and empirical articles. These books are in my opinion definitive classics. But they didn't satisfy my needs, and as a result, I was constantly jumping between material. Other books were awesome but not quite right for me either. Imbens and Rubin [2015] cover the potential outcomes model, experimental design, and matching and instrumental variables, but not directed acyclic graphical models (DAGs), regression discontinuity, panel data, or synthetic control. Morgan and Winship [2014] cover DAGs, the potential outcomes model, and instrumental variables, but have too light a touch on regression discontinuity and panel data for my tastes. They also don't cover synthetic control, which has been called the most important innovation in causal inference of the last 15 years by Athey and Imbens [2017b]. Angrist and Pischke [2009] is very close to what I need but does not include anything on synthetic control or on the graphical models that I find so critically useful. But maybe most importantly, Imbens and Rubin

[2015], Angrist and Pischke [2009], and Morgan and Winship [2014] do not provide *any* practical programming guidance, and I believe it is in replication and coding that we gain knowledge in these areas.[4]

This book was written with a few different people in mind. It was written first and foremost for *practitioners*, which is why it includes easy-to-download data sets and programs. It's why I have made several efforts to review papers as well as replicate the models as much as possible. I want readers to understand this field, but as important, I want them to feel empowered so that they can use these tools to answer their own research questions.

Another person I have in mind is the experienced social scientist who wants to retool. Maybe these are people with more of a theoretical bent or background, or maybe they're 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 estimation. The DAGs in particular are valuable for this group, I think.

A third group that I'm focusing on is the nonacademic person in industry, media, think tanks, and the like. Increasingly, knowledge about causal inference is expected throughout the professional world. It is no longer simply something that academics sit around and debate. It is crucial knowledge for making business decisions as well as for interpreting policy.

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

Do Not Confuse Correlation with Causality

It is very common these days to hear someone say "correlation does not mean causality." Part of the purpose of this book is to help readers be able to understand exactly why correlations, particularly

in observational data, are unlikely to be reflective of a causal relationship. When the rooster crows, the sun soon after rises, but we know the rooster didn't cause the sun to rise. Had the rooster been eaten by the farmer's cat, the sun still would have risen. Yet so often people make this kind of mistake when naively interpreting simple correlations.

But weirdly enough, sometimes there are causal relationships between two things and yet *no observable correlation*. Now that is definitely strange. How can one thing cause another thing without any discernible correlation between the two things? Consider this example, which is illustrated in Figure 1. A sailor is sailing her boat across the lake on a windy day. As the wind blows, she counters by turning the rudder in such a way so as to exactly offset the force of the wind. Back and forth she moves the rudder, yet the boat follows a straight line across the lake. A kindhearted yet naive person with no knowledge of wind or boats might look at this woman and say, "Someone get this sailor a new rudder! Hers is broken!" He thinks this because he cannot see any relationship between the movement of the rudder and the direction of the boat.

Figure 1. No correlation doesn't mean no causality. Artwork by Seth Hahne © 2020.

But does the fact that he cannot see the relationship mean there isn't one? Just because there is no observable relationship does not mean there is no causal one. Imagine that instead of perfectly countering the wind by turning the rudder, she had instead flipped a coin—heads she turns the rudder left, tails she turns the rudder right. What do you think this man would have seen if she was sailing her boat according to coin flips? If she *randomly* moved the rudder on a windy day, then he would see a sailor zigzagging across the lake. Why would he see the relationship if the movement were randomized but not be able to see it otherwise? Because the sailor is *endogenously* moving the rudder in response to the unobserved wind. And as such, the relationship between the rudder and the boat's direction is canceled—even though there is a causal relationship between the two.

This sounds like a silly example, but in fact there are more serious versions of it. Consider a central bank reading tea leaves to discern

when a recessionary wave is forming. Seeing evidence that a recession is emerging, the bank enters into open-market operations, buying bonds and pumping liquidity into the economy. Insofar as these actions are done optimally, these open-market operations will show no relationship whatsoever with actual output. In fact, in the ideal, banks may engage in aggressive trading in order to stop a recession, and we would be unable to see any evidence that it was working *even though it was*!

Human beings engaging in optimal behavior are the main reason correlations almost never reveal causal relationships, because rarely are human beings acting randomly. And as we will see, it is the presence of randomness that is crucial for identifying causal effect.

Optimization Makes Everything Endogenous

Certain presentations of causal inference methodologies have sometimes been described as atheoretical, but in my opinion, while some practitioners seem comfortable flying blind, the actual methods employed in causal designs are always deeply dependent on theory and local institutional knowledge. It is my firm belief, which I will emphasize over and over in this book, that without prior knowledge, estimated causal effects are rarely, if ever, believable. Prior knowledge is *required* in order to justify any claim of a causal finding. 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

Esther Duflo, Michael Kremer, Abhijit Banerjee, and John List, and many others.

Observational data is usually collected through surveys in a retrospective manner, or as the by-product of some other business activity ("big data"). In many observational studies, you collect data about what happened previously, as opposed to collecting data as it happens, though with the increased use of web scraping, it may be possible to get observational data closer to the exact moment in which some action occurred. But regardless of the timing, the researcher is a passive actor in the processes creating the data itself. She observes actions and results but is not in a position to interfere with the environment in which the units under consideration exist. This is the most common form of data that many of us will ever work with.

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. And we see this problem reflected in the potential outcomes model itself: a correlation, in order to be a measure of a causal effect, must be based on a choice that was made 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 is based on potential outcomes, and the correlation does not remotely satisfy the conditions we need in order to say it is causal. To put it as bluntly as I can, economic theory says choices are endogenous, and therefore since they are, the correlations between those choices and outcomes in the aggregate will rarely, if ever, represent a causal effect.

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, asterisks). The foundations of scientific knowledge are scientific methodologies. True scientists do not collect evidence in order to prove what they want to be true or what others want to believe. That is a form of deception and manipulation called *propaganda*, and propaganda is 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 believable.

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 and make falsifiable (testable) predictions. A prediction is falsifiable insofar as we can evaluate, and potentially reject, the prediction with data.[6] A 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 a model, we turn it into what is called an econometric model, which can be estimated 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 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 the idea of *ceteris paribus*—or "all else constant." When we are trying to describe the causal effect of some intervention, for instance, we are always assuming that the other relevant variables in the model are not changing. If they were changing, then they would be correlated with the variable of interest and it would confound our estimation.[8]

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 because knowing it enables firms to maximize profits and governments to choose optimal taxes, and whether to restrict quantity altogether [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 emerges in history that reflect neither the demand curve nor the supply curve. In fact, connecting the points does not reflect any meaningful or useful object.

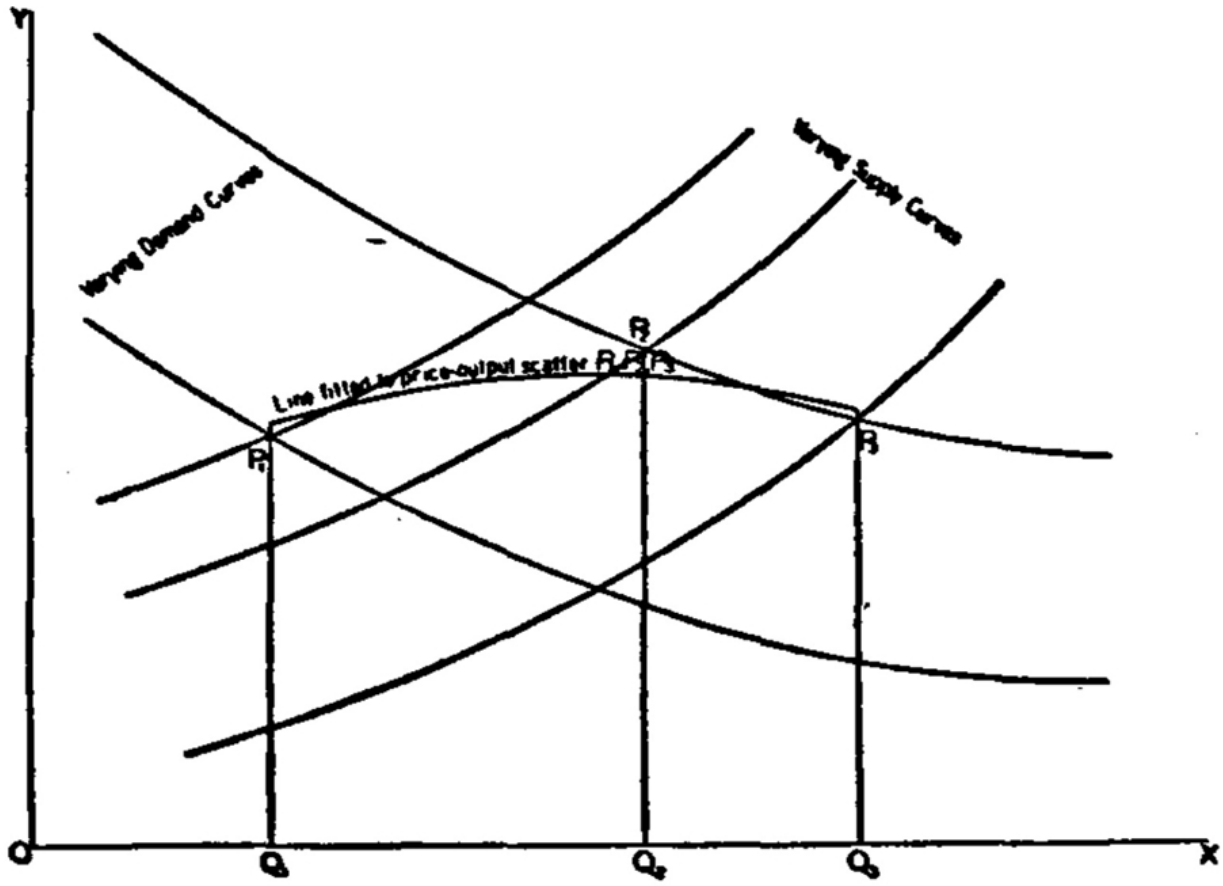


Figure 2. Wright's graphical demonstration of the identification problem. Figure from Wright, P. G. (1928). *The Tariff on Animal and Vegetable Oils*. The Macmillan Company.

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, and so on. 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 elasticity of demand.

The problem is that the elasticity is an important object, and we need to know it, and therefore we need to solve this problem. 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.[9]

Foreshadowing the content of this mixtape, we need two things to estimate price elasticity of demand. First, we need numerous rows of data on price and quantity. Second, we need for the variation in price in our imaginary data set 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 stabs in the dark.

Conclusion

This book is an introduction to research designs that can recover causal effects. But just as importantly, it provides you with hands-on practice to implement these designs. Implementing these designs means writing code in some type of software. I have chosen to illustrate these designs using two popular software languages: Stata (most commonly used by economists) and R (most commonly used by everyone else).

The book contains numerous empirical exercises illustrated in the Stata and R programs. These exercises are either simulations (which don't need external data) or exercises requiring external data. The data needed for the latter have been made available to you at Github. The Stata examples will download files usually at the start of

the program using the following command: use github.com/scunning1975/mixtape/raw/master/DATAFILENAME.DTA, where DATAFILENAME.DTA is the name of a particular data set.

For R users, it is a somewhat different process to load data into memory. In an effort to organize and clean the code, my students Hugo Sant'Anna and Terry Tsai created a function to simplify the data download process. This is partly based on a library called haven, which is a package for reading data files. It is secondly based on a set of commands that create a function that will then download the data directly from Github.[10]

Some readers may not be familiar with either Stata or R but nonetheless wish to follow along. I encourage you to use this opportunity to invest in learning one or both of these languages. It is beyond the scope of this book to provide an introduction to these languages, but fortunately, there are numerous resources online. For instance, Christopher Baum has written an excellent introduction to Stata at https://fmwww.bc.edu/GStat/docs/StataIntro.pdf. Stata is popular among microeconomists, and given the amount of coauthoring involved in modern economic research, an argument could be made for investing in it solely for its ability to solve basic coordination problems between you and potential coauthors. But a downside to Stata is that it is proprietary and must be purchased. And for some people, that may simply be too big of a barrier—especially for anyone simply wanting to follow along with the book. R on the other hand is open-source and free. Tutorials on Basic R can be found at https://cran.r-project.org/doc/contrib/Paradis-rdebuts_en.pdf, and an introduction to Tidyverse (which is used throughout the R programming) can be found at https://r4ds.had.co.nz. Using this time to learn R would likely be well worth your time.

Perhaps you already know R and want to learn Stata. Or perhaps you know Stata and want to learn R. Then this book may be helpful because of the way in which both sets of code are put in sequence to accomplish the same basic tasks. But, with that said, in many situations, although I have tried my best to reconcile results from

Stata and R, I was not always able to do so. Ultimately, Stata and R are different programming languages that sometimes yield different results because of different optimization procedures or simply because the programs are built slightly differently. This has been discussed occasionally in articles in which authors attempt to better understand what accounts for the differing results. I was not always able to fully reconcile different results, and so I offer the two programs as simply alternative approaches. You are ultimately responsible for anything you do on your own using either language for your research. I leave it to you ultimately to understand the method and estimating procedure contained within a given software and package.

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 the causal methodologies themselves, though, I need to lay down a foundation in statistics and regression modeling. Buckle up! This is going to be fun.

Notes

1 "Too long; didn't read."

2 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. Interestingly, when I later found economics, I went back to Rilke and asked myself if I could live without it. This time, I decided I couldn't, or wouldn't—I wasn't sure which. So I stuck with it and got a PhD.

3 I decided to write this book for one simple reason: I didn't feel that the market had provided the book that I needed for my students. So I wrote this book for my students and me so that we'd all be on the same page. This book is my best effort to explain causal inference to myself. I felt that if I could explain causal inference to myself, then I would be able to explain it to others too. Not thinking the book would have much value outside of my class, I posted it to my website and told people about it on Twitter. I was surprised to learn that so many people found the book helpful.

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

5 It is not the only cornerstone, or 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 through an intuitive and less formal reasoning process. But economics favors formalism and deductive methods.

7 Scientific models, be they economic ones or otherwise, are abstract, not realistic, representations of the world. That is a strength, not a weakness. George Box, the statistician, once quipped that "all models are wrong, but some are useful." A model's usefulness is its ability to unveil hidden secrets about the world. No more and no less.

8 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 the unobservable covariates.

9 More on the error term later.

10 This was done solely for aesthetic reasons. Often the URL was simply too long for the margins of the book otherwise.

Regression Discontinuity

Jump around!
Jump around!
Jump up, jump up, and get down!
Jump!
House of Pain

Huge Popularity of Regression Discontinuity

Waiting for life. Over the past twenty years, interest in the *regression-discontinuity design* (RDD) has increased (Figure 19). It was not always so popular, though. The method dates back about sixty years to Donald Campbell, an educational psychologist, who wrote several studies using it, beginning with Thistlehwaite and Campbell [1960].[1] In a wonderful article on the history of thought around RDD, Cook [2008] documents its social evolution. Despite Campbell's many efforts to advocate for its usefulness and understand its properties, RDD did not catch on beyond a few doctoral students and a handful of papers here and there. Eventually, Campbell too moved on from it.

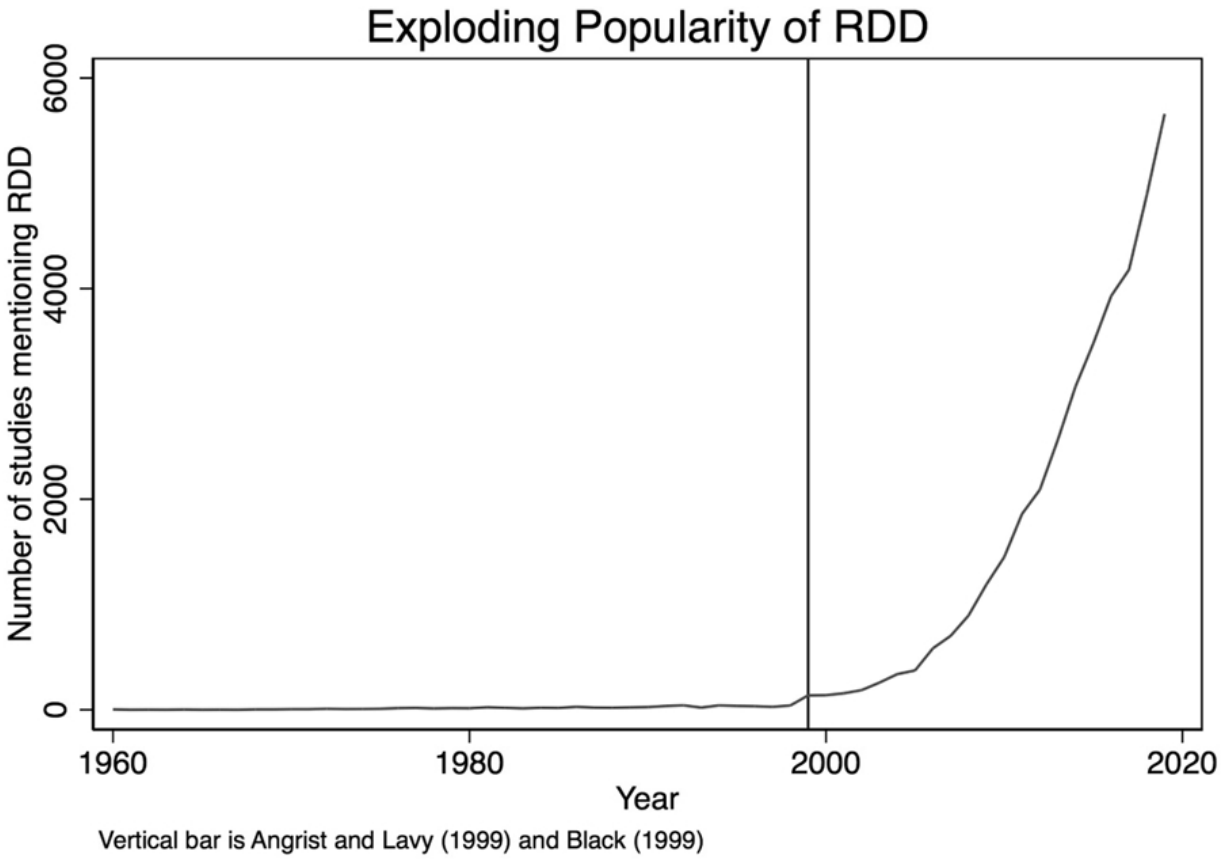


Figure 19. Regression discontinuity over time.

To see its growing popularity, let's look at counts of papers from Google Scholar by year that mentioned the phrase "regression discontinuity design" (see Figure 19).[2] Thistlehwaite and Campbell [1960] had no influence on the broader community of scholars using his design, confirming what Cook [2008] wrote. The first time RDD appears in the economics community is with an unpublished econometrics paper [Goldberger, 1972]. Starting in 1976, RDD finally gets annual double-digit usage for the first time, after which it begins to slowly tick upward. But for the most part, adoption was imperceptibly slow.

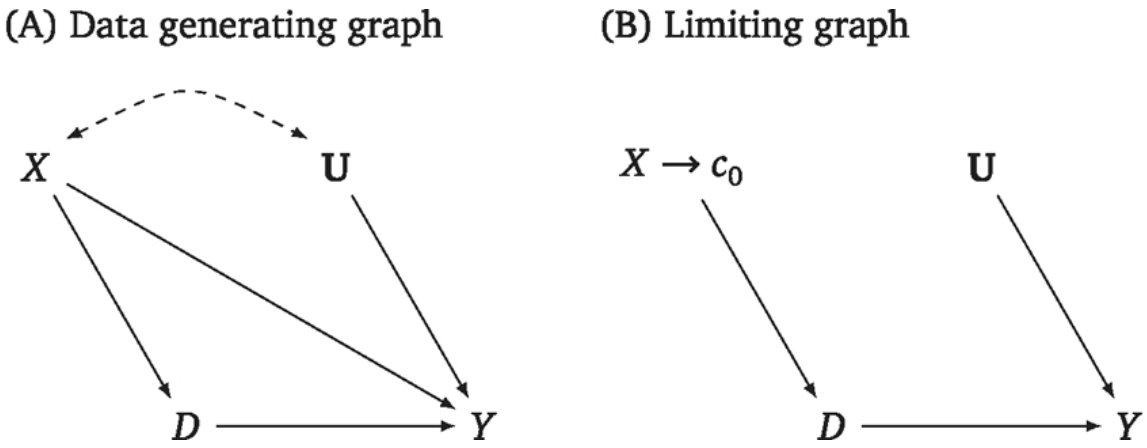
But then things change starting in 1999. That's the year when a couple of notable papers in the prestigious *Quarterly Journal of Economics* resurrected the method. These papers were Angrist and Lavy [1999] and Black [1999], followed by Hahn et al. [2001] two years later. Angrist and Lavy [1999], which we discuss in detail later, studied the effect of class size on pupil achievement using an

unusual feature in Israeli public schools that created smaller classes when the number of students passed 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. The year 1999 marks a watershed in the design's widespread adoption. A 2010 *Journal of Economic Literature* article by Lee and Lemieux, which has nearly 4,000 cites shows up in a year with nearly 1,500 new papers mentioning the method. By 2019, RDD output would be over 5,600. The design is today incredibly popular and shows no sign of slowing down.

But 1972 to 1999 is a long time without so much as a peep for what is now considered one of the most credible research designs with observational data, so what gives? Cook [2008] says that RDD was "waiting for life" during this time. The conditions for life in empirical microeconomics were likely the growing acceptance of the potential outcomes framework among microeconomists (i.e., the so-called credibility revolution led by Josh Angrist, David Card, Alan Krueger, Steven Levitt, and many others) as well as, and perhaps even more importantly, the increased availability of large digitized the administrative data sets, many of which often captured unusual administrative rules for treatment assignments. These unusual rules, combined with the administrative data sets' massive size, provided the much-needed necessary conditions for Campbell's original design to bloom into thousands of flowers.

Graphical representation of RDD. So what's the big deal? Why is RDD so special? The reason RDD is so appealing to many is because of its ability to convincingly eliminate selection bias. This appeal is partly due to the fact that its underlying identifying assumptions are viewed by many as easier to accept and evaluate. Rendering selection bias impotent, the procedure is capable of recovering average treatment effects for a given subpopulation of units. The method is based on a simple, intuitive idea. Consider the following DAG developed by Steiner et al. [2017] that illustrates this method very well.

In the first graph, X is a continuous variable assigning units to treatment D ($X \rightarrow D$). This assignment of units to treatment is based on a "cutoff" score c_0 such that any unit with a score above the cutoff gets placed into the treatment group, and units below do not. An example might be a charge of driving while intoxicated (or impaired; DWI). Individuals with a blood-alcohol content of 0.08 or higher are arrested and charged with DWI, whereas those with a blood-alcohol level below 0.08 are not [Hansen, 2015]. The assignment variable may itself independently affect the outcome via the $X \rightarrow Y$ path and may even be related to a set of variables U that independently determine Y. Notice for the moment that a unit's treatment status is *exclusively* determined by the assignment rule. Treatment is not determined by U.



This DAG clearly shows that the assignment variable X—or what is often called the "running variable"—is an observable confounder since it causes both D and Y. Furthermore, because the assignment variable assigns treatment on the basis of a cutoff, we are never able to observe units in both treatment and control for the same value of X. Calling back to our matching chapter, this means a situation such as this one does not satisfy the overlap condition needed to use matching methods, and therefore the backdoor criterion cannot be met.[3]

However, we can identify causal effects using RDD, which is illustrated in the limiting graph DAG. We can identify causal effects

for those subjects whose score is in a close neighborhood around some cutoff c_0. Specifically, as we will show, the average causal effect for this subpopulation is identified as $X \rightarrow c_0$ in the limit. This is possible because the cutoff is the sole point where treatment and control subjects overlap in the limit.

There are a variety of explicit assumptions buried in this graph that must hold in order for the methods we will review later to recover any average causal effect. But the main one I discuss here is that the cutoff itself cannot be endogenous to some competing intervention occurring at precisely the same moment that the cutoff is triggering units into the D treatment category. This assumption is called *continuity*, and what it formally means is that the expected potential outcomes are continuous at the cutoff. If expected potential outcomes are continuous at the cutoff, then it necessarily rules out competing interventions occurring at the same time.

The continuity assumption is reflected graphically by the absence of an arrow from $X \rightarrow Y$ in the second graph because the cutoff c_0 has cut it off. At c_0, the assignment variable X no longer has a direct effect on Y. Understanding continuity should be one of your main goals in this chapter. It is my personal opinion that the null hypothesis should always be continuity and that any discontinuity necessarily implies some cause, because the tendency for things to change gradually is what we have come to expect in nature. Jumps are so unnatural that when we see them happen, they beg for explanation. Charles Darwin, in his *On the Origin of Species*, summarized this by saying *Natura non facit saltum*, or "nature does not make jumps." Or to use a favorite phrase of mine from growing up in Mississippi, if you see a turtle on a fencepost, you know he didn't get there by himself.

That's the heart and soul of RDD. We use our knowledge about selection into treatment in order to estimate average treatment effects. Since we know the probability of treatment assignment changes discontinuously at c_0, then our job is simply to compare people above and below c_0 to estimate a particular kind of average

treatment effect called the *local average treatment effect*, or LATE [Imbens and Angrist, 1994]. Because we do not have overlap, or "common support," we must rely on extrapolation, which means we are comparing units with different values of the running variable. They only overlap in the limit as *X* approaches the cutoff from either direction. All methods used for RDD areways of handling the bias from extrapolation as cleanly as possible. *A picture is worth a thousand words*. As I've said before, and will say again and again—pictures of your main results, including your identification strategy, are absolutely essential to any study attempting to convince readers of a causal effect. And RDD is no different. In fact, pictures are the comparative advantage of RDD. RDD is, like many modern designs, a very visually intensive design. It and synthetic control are probably two of the most visually intensive designs you'll ever encounter, in fact. So to help make RDD concrete, let's first look at a couple of pictures. The following discussion derives from Hoekstra [2009].[4]

Labor economists had for decades been interested in estimating the causal effect of college on earnings. But Hoekstra wanted to crack open the black box of college's returns a little by checking whether there were heterogeneous returns to college. He does this by estimating the causal effect of attending the state flagship university on earnings. State flagship universities are often more selective than other public universities in the same state. In Texas, the top 7% of graduating high school students can select their university in state, and the modal first choice is University of Texas at Austin. These universities are often environments of higher research, with more resources and strongly positive peer effects. So it is natural to wonder whether there are heterogeneous returns across public universities.

The challenge in this type of question should be easy to see. Let's say that we were to compare individuals who attended the University of Florida to those who attended the University of South Florida. Insofar as there is positive selection into the state flagship school, we might expect individuals with higher observed and unobserved ability to sort into the state flagship school. And insofar as that ability

increases one's marginal product, then we expect those individuals to earn more in the workforce regardless of whether they had in fact attended the state flagship. Such basic forms of selection bias confound our ability to estimate the causal effect of attending the state flagship on earnings. But Hoekstra [2009] had an ingenious strategy to disentangle the causal effect from the selection bias using an RDD. To illustrate, let's look at two pictures associated with this interesting study.

Before talking about the picture, I want to say something about the data. Hoekstra has data on all applications to the state flagship university. To get these data, he would've had to build a relationship with the admissions office. This would have involved making introductions, holding meetings to explain his project, convincing administrators the project had value for them as well as him, and ultimately winning their approval to cooperatively share the data. This likely would've involved the school's general counsel, careful plans to de-identify the data, agreements on data storage, and many other assurances that students' names and identities were never released and could not be identified. There is a lot of trust and social capital that must be created to do projects like this, and this is the secret sauce inmost RDDs—your acquisition of the data requires far more soft skills, such as friendship, respect, and the building of alliances, than you may be accustomed to. This isn't as straightforward as simply downloading the CPS from IPUMS; it's going to take genuine smiles, hustle, and luck. Given that these agencies have considerable discretion in whom they release data to, it is likely that certain groups will have more trouble than others in acquiring the data. So it is of utmost importance that you approach these individuals with humility, genuine curiosity, and most of all, scientific integrity. They ultimately are the ones who can give you the data if it is not public use, so don't be a jerk.[5]

But on to the picture. Figure 20 has a lot going on, and it's worth carefully unpacking each element for the reader. There are four distinct elements to this picture that I want to focus on. First, notice the horizontal axis. It ranges from a negative number to a positive

number with a zero around the center of the picture. The caption reads “SAT Points Above or Below the Admission Cutoff.” Hoekstra has “recentered” the university’s admissions criteria by subtracting the admission cutoff from the students’ actual score, which is something I discuss in more detail later in this chapter. The vertical line at zero marks the “cutoff,” which was this university’s minimum SAT score for admissions. It appears it was binding, but not deterministically, for there are some students who enrolled but did not have the minimum SAT requirements. These individuals likely had other qualifications that compensated for their lower SAT scores. This recentered SAT score is in today’s parlance called the “running variable.”

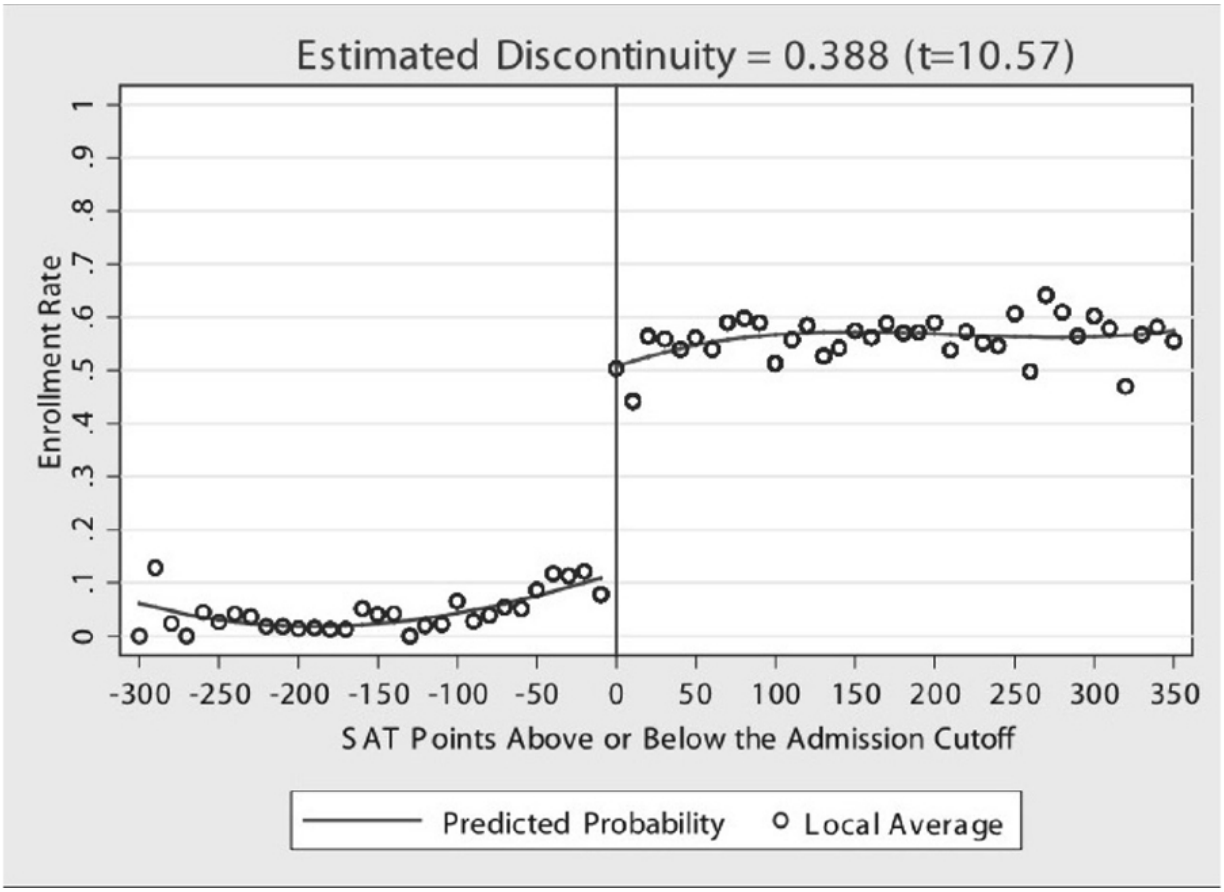


Figure 20. Attending the state flagship university as a function of recentered standardized test scores. Reprinted from Mark Hoekstra, "The Effect of Attending the Flagship State University on Earnings: A Discontinuity-Based Approach," *The Review of Economics and Statistics*, 91:4 (November, 2009), pp. 717–724. © 2009 by the President and Fellows of Harvard College and the Massachusetts Institute of Technology.

Second, notice the dots. Hoekstra used hollow dots at regular intervals along the recentered SAT variable. These dots represent conditional mean enrollments per recentered SAT score. While his administrative data set contains thousands and thousands of observations, he only shows the conditional means along evenly spaced out bins of the recentered SAT score.

Third are the curvy lines fitting the data. Notice that the picture has *two* such lines—there is a curvy line fitted to the left of zero, and there is a separate line fit to the right. These lines are the least squares fitted values of the running variable, where the running variable was allowed to take on higher-order terms. By including

higher-order terms in the regression itself, the fitted values are allowed to more flexibly track the central tendencies of the data itself. But the thing I really want to focus your attention on is that there are two lines, not one. He fit the lines separately to the left and right of the cutoff.

Finally, and probably the most vivid piece of information in this picture—the gigantic jump in the dots at zero on the recentered running variable. What is going on here? Well, I think you probably know, but let me spell it out. The probability of enrolling at the flagship state university jumps discontinuously when the student just barely hits the minimum SAT score required by the school. Let's say that the score was 1250. That means a student with 1240 had a lower chance of getting in than a student with 1250. Ten measly points and they have to go a different path.

Imagine two students—the first student got a 1240, and the second got a 1250. Are these two students really so different from one another? Well, sure: those two *individual* students are likely very different. But what if we had hundreds of students who made 1240 and hundreds more who made 1250. Don't you think those two groups are probably pretty similar to one another on observable and unobservable characteristics? After all, why would there be suddenly at 1250 a major difference in the characteristics of the students in a large sample? That's the question you should reflect on. If the university is arbitrarily picking a reasonable cutoff, are there reasons to believe they are also picking a cutoff where the natural ability of students jumps at that exact spot?

But I said Hoekstra is evaluating the effect of attending the state flagship university on future earnings. Here's where the study gets even more intriguing. States collect data on workers in a variety of ways. One is through unemployment insurance tax reports. Hoekstra's partner, the state flagship university, sent the university admissions data directly to a state office in which employers submit unemployment insurance tax reports. The university had social security numbers, so the matching of student to future worker worked quite well since a social security number uniquely identifies a worker. The social security numbers were used to match quarterly

earnings records from 1998 through the second quarter of 2005 to the university records. He then estimated:

$$\ln(\text{Earnings}) = \psi_{\text{Year}} + \omega_{\text{Experience}} + \theta_{\text{Cohort}} + \varepsilon$$

where ψ is a vector of year dummies, ω is a dummy for years after high school that earnings were observed, and θ is a vector of dummies controlling for the cohort in which the student applied to the university (e.g., 1988). The residuals from this regression were then averaged for each applicant, with the resulting average residual earnings measure being used to implement a partialled out future earnings variable according to the Frisch-Waugh-Lovell theorem. Hoekstra then takes each students' residuals from the natural log of earnings regression and collapses them into conditional averages for bins along the recentered running variable. Let's look at that in Figure 21.

In this picture, we see many of the same elements we saw in Figure 20. For instance, we see the recentered running variable along the horizontal axis, the little hollow dots representing conditional means, the curvy lines which were fit left and right of the cutoff at zero, and a helpful vertical line at zero. But now we also have an interesting title: "Estimated Discontinuity = 0.095 (z = 3.01)." What is this exactly?

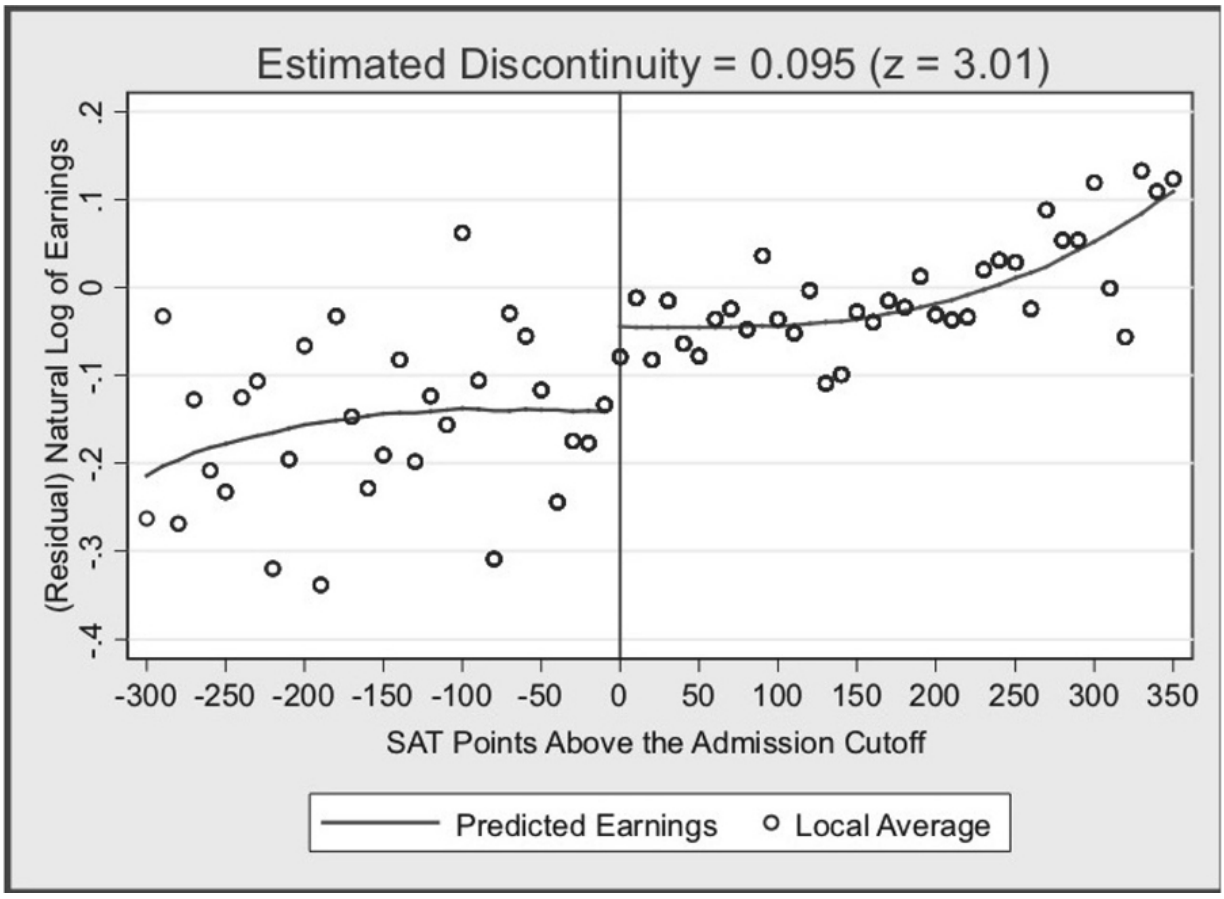


Figure 21. Future earnings as a function of recentered standardized test scores. Reprinted from Mark Hoekstra, "The Effect of Attending the Flagship State University on Earnings: A Discontinuity-Based Approach," *The Review of Economics and Statistics*, 91:4 (November, 2009), pp. 717–724. © 2009 by the President and Fellows of Harvard College and the Massachusetts Institute of Technology.

The visualization of a discontinuous jump at zero in earnings isn't as compelling as the prior figure, so Hoekstra conducts hypothesis tests to determine if the mean between the groups just below and just above are the same. He finds that they are not: those just above the cutoff earn 9.5% higher wages in the long term than do those just below. In his paper, he experiments with a variety of binning of the data (what he calls the "bandwidth"), and his estimates when he does so range from 7.4% to 11.1%.

Now let's think for a second about what Hoekstra is finding. Hoekstra is finding that at exactly the point where workers experienced a jump in the probability of enrolling at the state flagship

university, there is, ten to fifteen years later, a separate jump in logged earnings of around 10%. Those individuals who just barely made it in to the state flagship university made around 10% more in long-term earnings than those individuals who just barely missed the cutoff.

This, again, is the heart and soul of the RDD. By exploiting institutional knowledge about how students were accepted (and subsequently enrolled) into the state flagship university, Hoekstra was able to craft an ingenious natural experiment. And insofar as the two groups of applicants right around the cutoff have comparable future earnings in a world where neither attended the state flagship university, then there is no selection bias confounding his comparison. And we see this result in powerful, yet simple graphs. This study was an early one to show that not only does college matter for long-term earnings, but the sort of college you attend—even among public universities—matters as well.

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. And sometimes, if we are lucky, someone gives us the data that allows us to use these rules for our study.

I am convinced that firms and government agencies are unknowingly sitting atop a mountain of potential RDD-based projects. Students looking for thesis and dissertation ideas might try to find them. I encourage you to find a topic you are interested in and begin building relationships with local employers and government administrators for whom that topic is a priority. Take them out for coffee, get to know them, learn about their job, and ask them how treatment assignment works. Pay close attention to precisely how individual units get assigned to the program. Is it random? Is it via a rule? Oftentimes they will describe a process whereby a running variable is used for treatment assignment, but they won't call it that. While I can't promise this will yield pay dirt, my hunch, based in part on experience, is that they will end up describing to you some

running variable that when it exceeds a threshold, people switch into some intervention. Building alliances with local firms and agencies can pay when trying to find good research ideas.

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 include the probability of being arrested for DWI jumping at greater than 0.08 blood-alcohol content [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; Barreca et al., 2011]; the probability of attending summer school when grades fall below some minimum level [Jacob and Lefgen, 2004], and as we just saw, the probability of attending the state flagship university jumping when the applicant's test scores exceed some minimum requirement [Hoekstra, 2009].

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 data sets 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 in the running variable, sample-size requirements get even larger. Researchers are typically using administrative data or settings such as birth records where there are many observations.

Estimation Using an RDD

The Sharp RD Design. There are generally accepted two kinds of RDD studies. There are designs where the probability of treatment goes from0 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 receiving some treatment flips.

Let's look at the diagram in Figure 22, which illustrates the similarities and differences between the two designs.

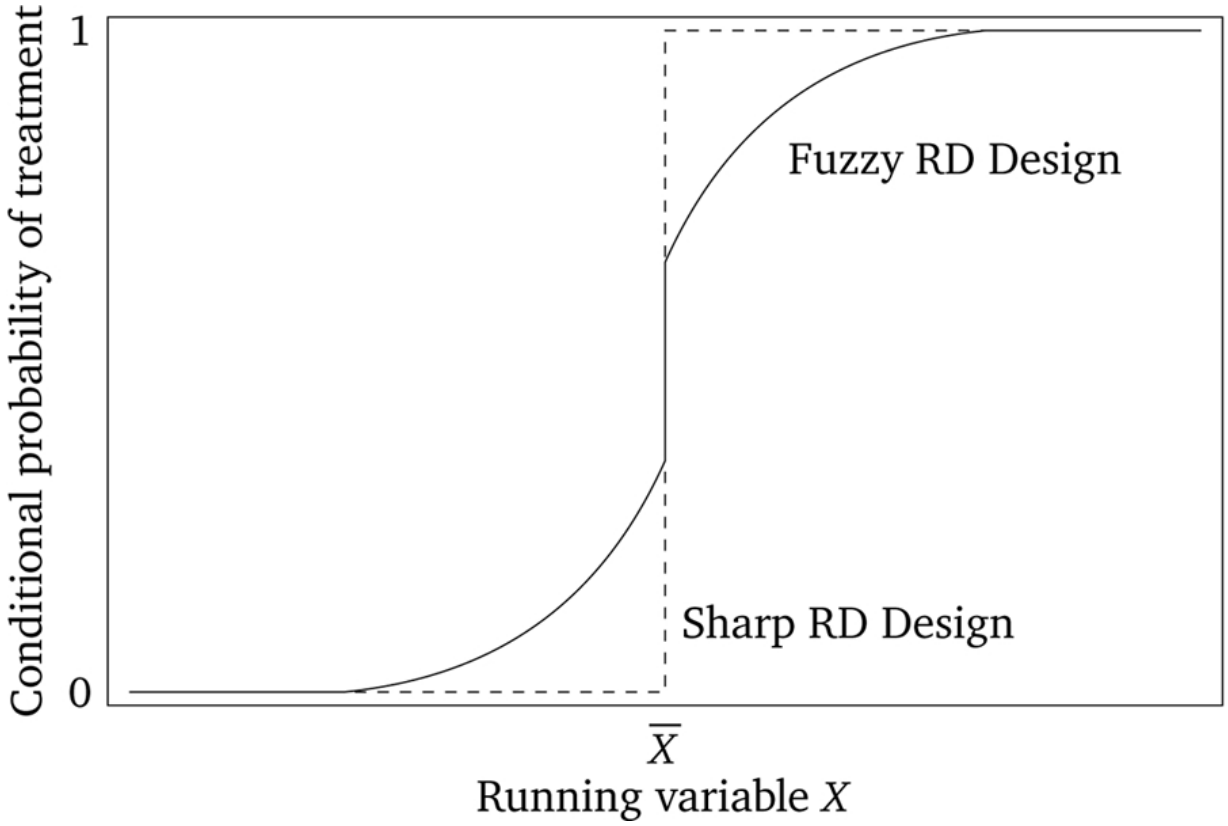


Figure 22. Sharp vs. Fuzzy RDD.

Sharp RDD is where treatment is a deterministic function of the running variable X.[6] An example might be Medicare enrollment, which happens sharply at age 65, excluding 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 a class-size function they created from the rules used by Israeli schools to construct class sizes.

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. If you know the value of X_i for unit i, then you know treatment assignment for unit i with certainty. But, if for every value of X you can perfectly predict the treatment assignment, then it necessarily means that there are no overlap along the running variable.

If we assume constant treatment effects, then in potential outcomes terms, we get

$$Y_i^0 = \alpha + \beta X_i$$
$$Y_i^1 = Y_i^0 + \delta$$

Using the switching equation, we get

$$Y_i = Y_i^0 + (Y_i^1 - Y_i^0)D_i$$
$$Y_i = \alpha + \beta X_i + \delta D_i + \varepsilon_i$$

where the treatment effect parameter, δ, is the discontinuity in the conditional expectation function:

$$\delta = \lim_{X_i \to X_0} E[Y_i^1 \mid X_i = X_0] - \lim_{X_0 \leftarrow X_i} E[Y_i^0 \mid X_i = X_0] \tag{6.1}$$

$$= \lim_{X_i \to X_0} E[Y_i \mid X_i = X_0] - \lim_{X_0 \leftarrow X_i} E[Y_i \mid X_i = X_0] \tag{6.2}$$

The sharp RDD estimation is interpreted as an average causal effect of the treatment as the running variable approaches the cutoff in the limit, for it is only in the limit that we have overlap. This average causal effect is the local average treatment effect (LATE). We discuss LATE in greater detail in the instrumental variables, but I will say one thing about it here. Since identification in an RDD is a limiting case, we are technically only identifying an average causal effect for those units at the cutoff. Insofar as those units have

treatment effects that differ from units along the rest of the running variable, then we have only estimated an average treatment effect that is local to the range around the cutoff. We define this local average treatment effect as follows:

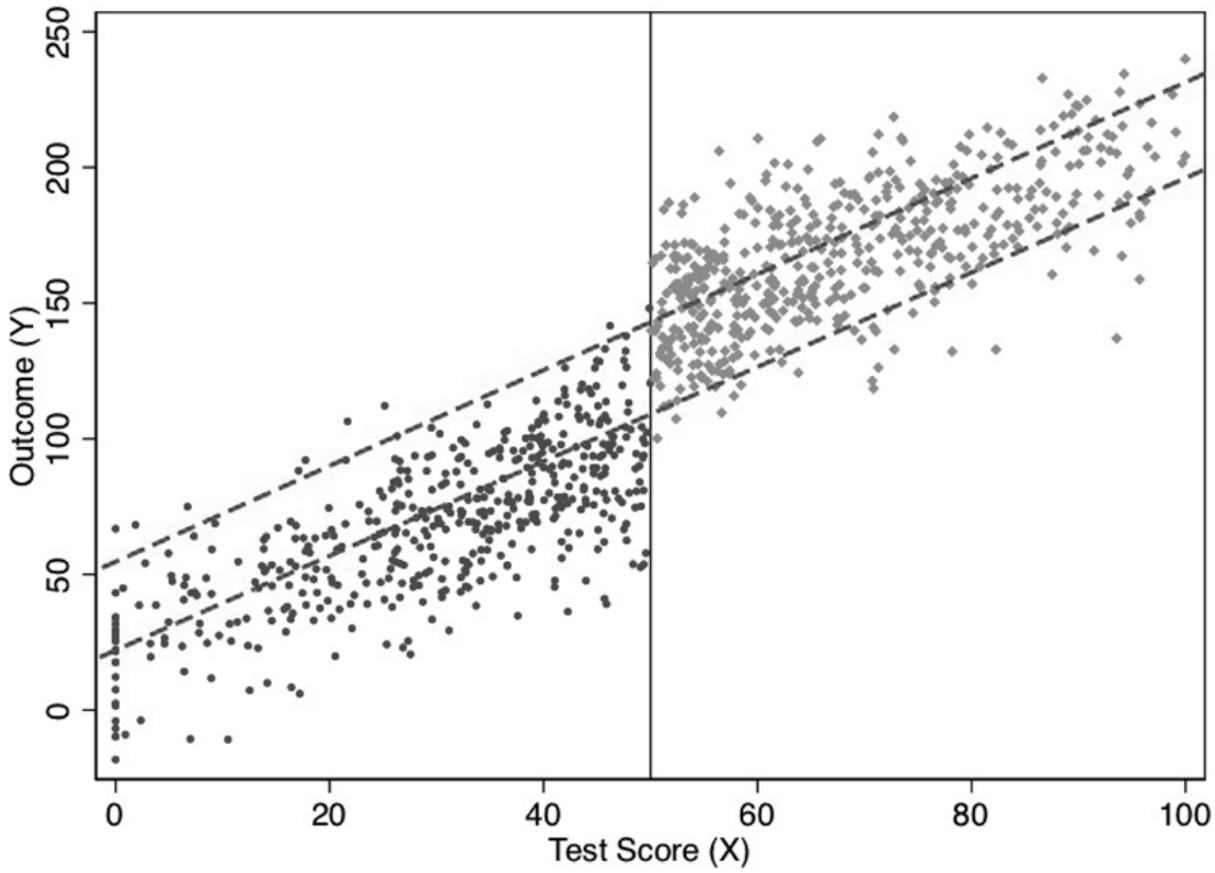


Figure 23. Simulated data representing observed data points along a running variable below and above some binding cutoff. *Note*: Dashed lines are extrapolations.

$$\delta_{SRD} = E[Y_i^1 - Y_i^0 \mid X_i = c_0] \tag{6.3}$$

Notice the role that *extrapolation* plays in estimating treatment effects with sharp RDD. If unit *i* is just below c_0, then D_i = 0. But if unit *i* is just above c_0, then the D_i =1. But for any value of X_i, there are either units in the treatment group or the control group, but not both. Therefore, the RDD does not have common support, which is one of the reasons we rely on extrapolation for our estimation. See Figure 23.

Continuity assumption. The key identifying assumption in an RDD is called the continuity assumption. It states that $E[Y_i^0 \mid X = c_0]$ and $E[Y_i^1 \mid X = c_0]$ are continuous (smooth) functions of X even across the c_0 threshold. Absent the treatment, in other words, the expected potential outcomes wouldn't have jumped; they would've remained smooth functions of X. But think about what that means for a moment. If the expected potential outcomes are not jumping at c_0, then there necessarily are no competing interventions occurring at c_0. Continuity, in other words, explicitly rules out omitted variable bias at the cutoff itself. All other unobserved determinants of Y are continuously related to the running variable X. 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.

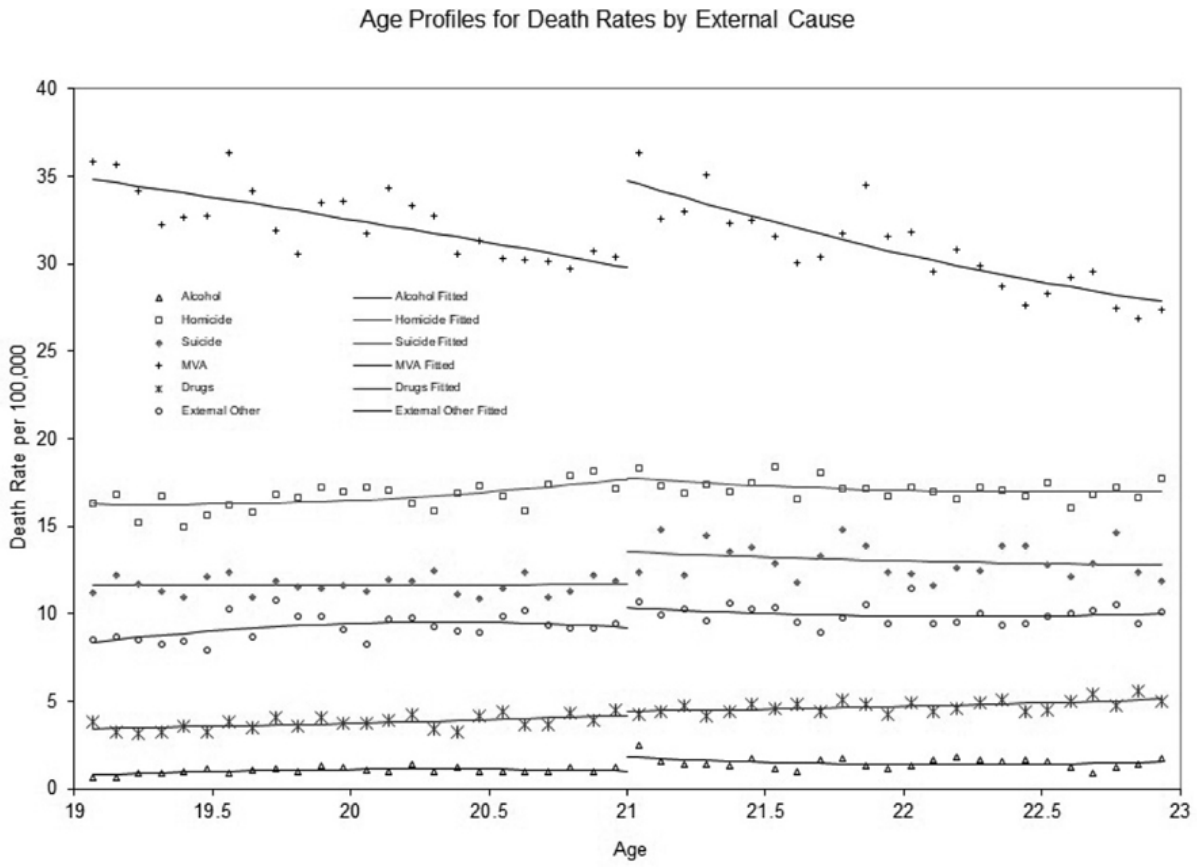


Figure 24. Mortality rates along age running variable [Carpenter and Dobkin, 2009].

I apologize if I'm beating a dead horse, but continuity is a subtle assumption and merits a little more discussion. The continuity assumption means that $E[Y^1 \mid X]$ wouldn't have jumped at c_0. If it had jumped, then it means something other than the treatment caused it to jump because Y^1 is already under treatment. So an example might be a study finding a large increase in motor vehicle accidents at age 21. I've reproduced a figure from and interesting study on mortality rates for different types of causes [Carpenter and Dobkin, 2009]. I have reproduced one of the key figures in Figure 24. Notice the large discontinuous jump in motor vehicle death rates at age 21. The most likely explanation is that age 21 causes people to drink more, and sometimes even while they are driving.

But this is only a causal effect if motor vehicle accidents don't jump at age 21 for other reasons. Formally, this is *exactly* what is implied

by continuity—the absence of simultaneous treatments at the cutoff. For instance, perhaps there is something biological that happens to 21-year-olds that causes them to suddenly become bad drivers. Or maybe 21-year-olds are all graduating from college at age 21, and during celebrations, they get into wrecks. To test this, we might replicate Carpenter and Dobkin [2009] using data from Uruguay, where the drinking age is 18. If we saw a jump in motor vehicle accidents at age 21 in Uruguay, then we might have reason to believe the continuity assumption does not hold in the United States. Reasonably defined placebos can help make the case that the continuity assumption holds, even if it is not a direct test per se.

Sometimes these abstract ideas become much easier to understand with data, so here is an example of what we mean using a simulation.

STATA

rdd_simulate1.do

```
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)

* Potential outcome Y1 not jumping at cutoff (continuity)
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)

```

R

rdd_simulate1.R

```
1  library(tidyverse)
2
3  # simulate the data
4  dat <- tibble(
5    x = rnorm(1000, 50, 25)
6  ) %>%
7    mutate(
8      x = if_else(x < 0, 0, x)
9    ) %>%
10   filter(x < 100)
11
12 # cutoff at x = 50
13 dat <- dat %>%
14   mutate(
15     D  = if_else(x > 50, 1, 0),
16     y1 = 25 + 0 * D + 1.5 * x + rnorm(n(), 0, 20)
17   )
18
19 ggplot(aes(x, y1, colour = factor(D)), data = dat) +
20   geom_point(alpha = 0.5) +
21   geom_vline(xintercept = 50, colour = "grey", linetype = 2)+
22   stat_smooth(method = "lm", se = F) +
23   labs(x = "Test score (X)", y = "Potential Outcome (Y1)")
```

Figure 25 shows the results from this simulation. Notice that the value of $E[Y^1 \mid X]$ is changing continuously over X and through c_0. This is an example of the continuity assumption. It means *absent the treatment itself*, the expected potential outcomes would've remained a smooth function of X even as passing c_0. Therefore, if continuity held, then *only* the treatment, triggered at c_0, could be responsible for discrete jumps in $E[Y \mid X]$.

The nice thing about simulations is that we actually observe the potential outcomes *since we made them ourselves*. But in the real world, we don't have data on potential outcomes. If we did, we could test the continuity assumption directly. But remember—by the

switching equation, we only observe actual outcomes, never potential outcomes. Thus, since units switch from Y^0 to Y^1 at c_0, we actually can't directly evaluate the continuity assumption. This is where institutional knowledge goes a long way, because it can help build the case that nothing else is changing at the cutoff that would otherwise shift potential outcomes.

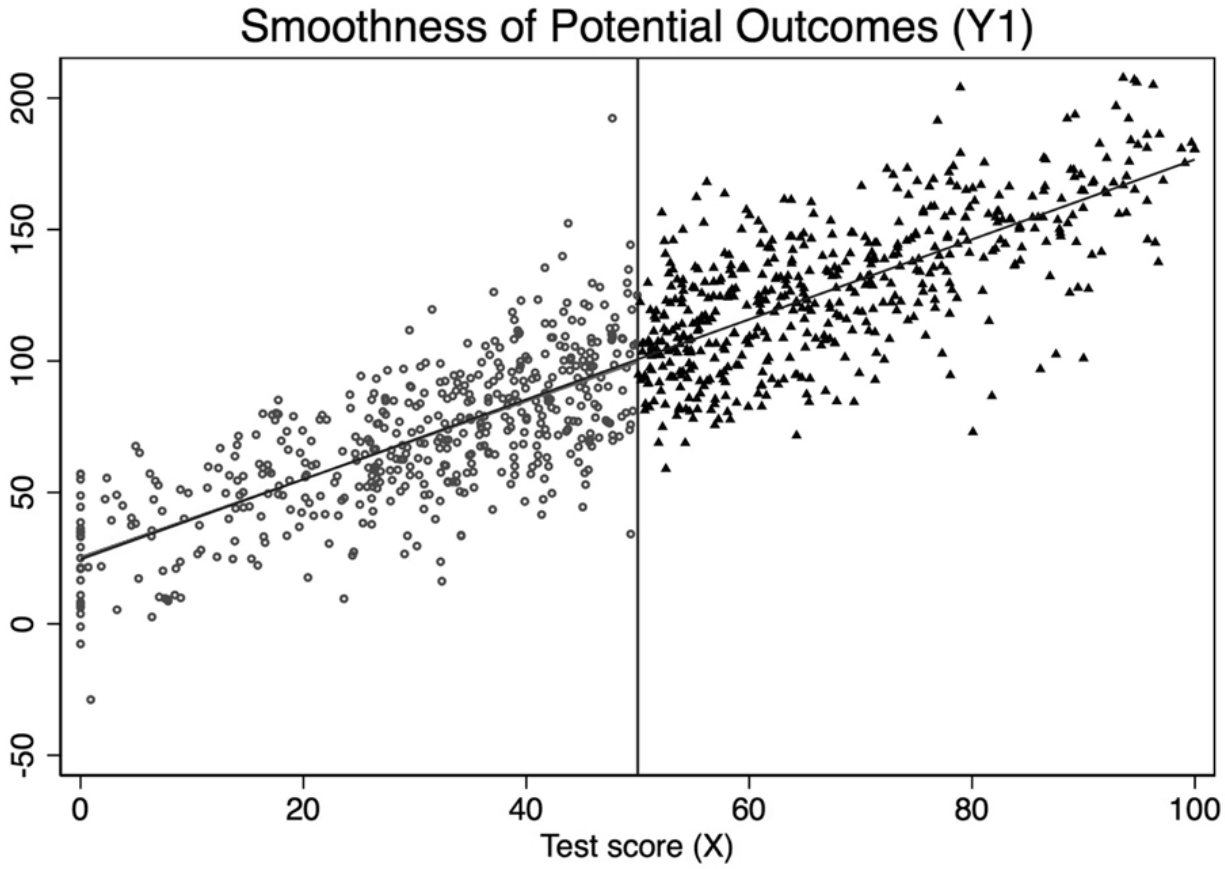


Figure 25. Smoothness of Y^1 across the cutoff illustrated using simulated data.

Let's illustrate this using simulated data. Notice that while Y^1 by construction had not jumped at 50 on the X running variable, Y will. Let's look at the output in Figure 26. Notice the jump at the discontinuity in the outcome, which I've labeled the LATE, or local average treatment effect.

STATA

rdd_simulate2.do

```
* Actual outcome jumping
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)")

```

R

rdd_simulate2.R

```
# simulate the discontinuity
dat <- dat %>%
  mutate(
    y2 = 25 + 40 * D + 1.5 * x + rnorm(n(), 0, 20)
  )

# figure 36
ggplot(aes(x, y2, colour = factor(D)), data = dat) +
  geom_point(alpha = 0.5) +
  geom_vline(xintercept = 50, colour = "grey", linetype = 2) +
  stat_smooth(method = "lm", se = F) +
  labs(x = "Test score (X)", y = "Potential Outcome (Y)")
```

Estimation using local and global least squares regressions. I'd like to now dig into the actual regression model you would use to estimate the LATE parameter in an RDD. We will first discuss some basic modeling choices that researchers often make—some trivial, some important. This section will focus primarily on regression-based estimation.

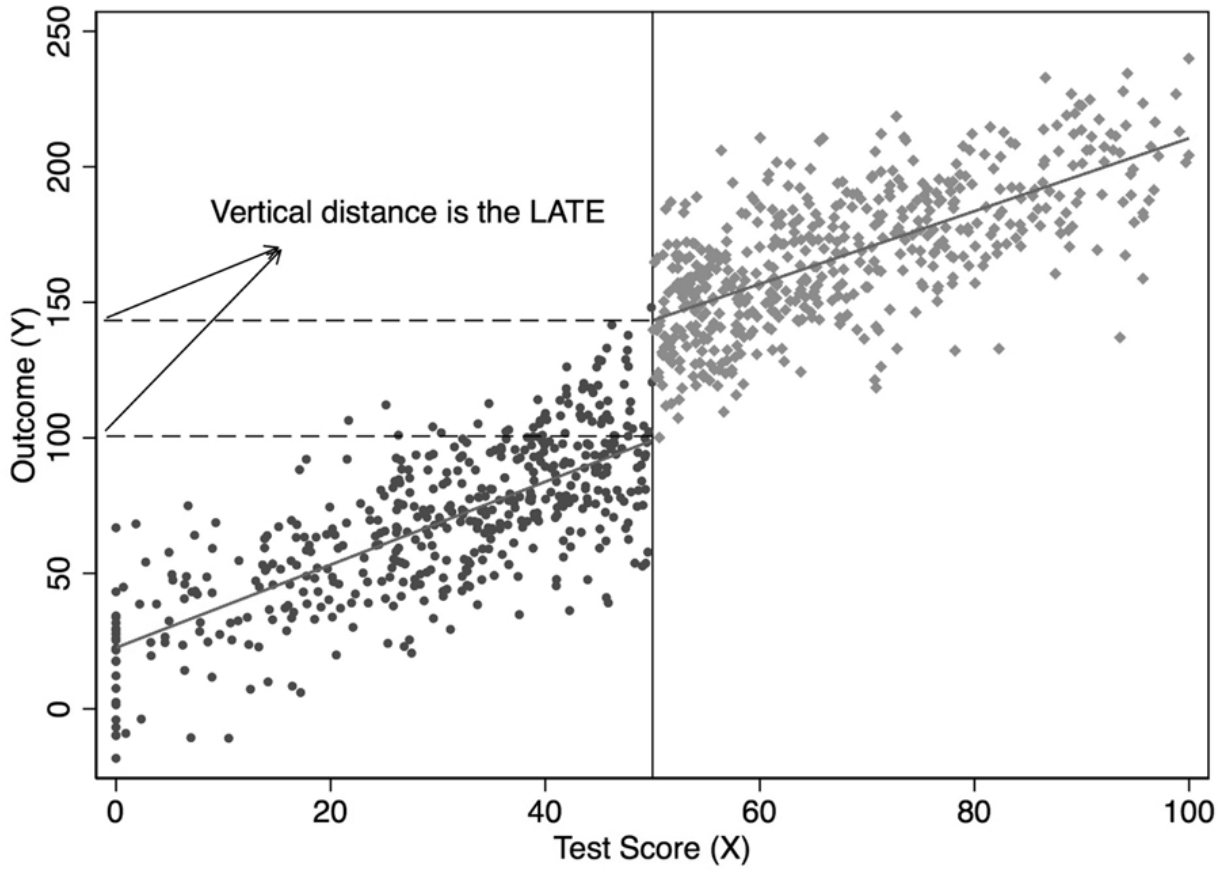


Figure 26. Estimated LATE using simulated data.

While not necessary, it is nonetheless quite common for authors to transform the running variable X by recentering 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 recenter the running variable (age) by subtracting 65:

$$\begin{aligned} Y &= \beta_0 + \beta_1(Age - 65) + \beta_2 Edu + \varepsilon \\ &= \beta_0 + \beta_1 Age - \beta_1 65 + \beta_2 Edu + \varepsilon \\ &= (\beta_0 - \beta_1 65) + \beta_1 Age + \beta_2 Edu + \varepsilon \\ &= \alpha + \beta_1 Age + \beta_2 Edu + \varepsilon \end{aligned}$$

where $\alpha = \beta_0 + \beta_1 65$. All other coefficients, notice, have the same interpretation except for the intercept.

Another practical question relates to nonlinear data-generating processes. A nonlinear data-generating process could easily yield false positives if we do not handle the specification carefully. Because sometimes we are fitting local linear regressions around the cutoff, we could spuriously pick up an effect simply for no other reason than that we imposed linearity on the model. But if the underlying data-generating process is nonlinear, then it may be a spurious result due to misspecification of the model. Consider an example of this nonlinearity in Figure 27.

STATA

rdd_simulate3.do

```
1  * Nonlinear data generating process
2  drop y y1 x* D
3  set obs 1000
4  gen x = rnormal(100, 50)
5  replace x=0 if x < 0
6  drop if x > 280
7  sum x, det
8
9  * Set the cutoff at X=140. Treated if X > 140
10 gen D = 0
```

(continued)

STATA *(continued)*

```
replace D = 1 if x > 140
gen x2 = x*x
gen x3 = x*x*x
gen y = 10000 + 0*D - 100*x +x2 + rnormal(0, 1000)
reg y D x

scatter y x if D==0, msize(vsmall) || scatter y x ///
  if D==1, msize(vsmall) legend(off) xline(140, ///
  lstyle(foreground)) ylabel(none) || lfit y x ///
  if D ==0, color(red) || lfit y x if D ==1, ///
  color(red) xtitle("Test Score (X)") ///
  ytitle("Outcome (Y)")

* Polynomial estimation
capture drop y
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)")
```

R

rdd_simulate3.R

```
# simultate nonlinearity
dat <- tibble(
   x = rnorm(1000, 100, 50)
 ) %>%
 mutate(
   x = case_when(x < 0 ~ 0, TRUE ~ x),
   D = case_when(x > 140 ~ 1, TRUE ~ 0),
   x2 = x*x,
   x3 = x*x*x,
   y3 = 10000 + 0 * D - 100 * x + x2 + rnorm(1000, 0, 1000)
 ) %>%
 filter(x < 280)
```

(continued)

R *(continued)*

```


ggplot(aes(x, y3, colour = factor(D)), data = dat) +
  geom_point(alpha = 0.2) +
  geom_vline(xintercept = 140, colour = "grey", linetype = 2) +
  stat_smooth(method = "lm", se = F) +
  labs(x = "Test score (X)", y = "Potential Outcome (Y)")

ggplot(aes(x, y3, colour = factor(D)), data = dat) +
  geom_point(alpha = 0.2) +
  geom_vline(xintercept = 140, colour = "grey", linetype = 2) +
  stat_smooth(method = "loess", se = F) +
  labs(x = "Test score (X)", y = "Potential Outcome (Y)")

```

I show this both visually and with a regression. As you can see in Figure 27, the data-generating process was nonlinear, but when with straight lines to the left and right of the cutoff, the trends in the running variable generate a spurious discontinuity at the cutoff. This shows up in a regression as well. When we fit the model using a least squares regression controlling for the running variable, we estimate a causal effect though there isn't one. In Table 40, the estimated effect of *D* on *Y* is large and highly significant, even though the true effect is zero. 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

$$E[Y_i^0 \mid X_i] = f(X_i)$$

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 $f(X_i)$. The traditional approaches let $f(X_B ei)$ equal a *pth*-order polynomial:

$$Y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_p x_i^p + \delta D_i + \eta_i$$

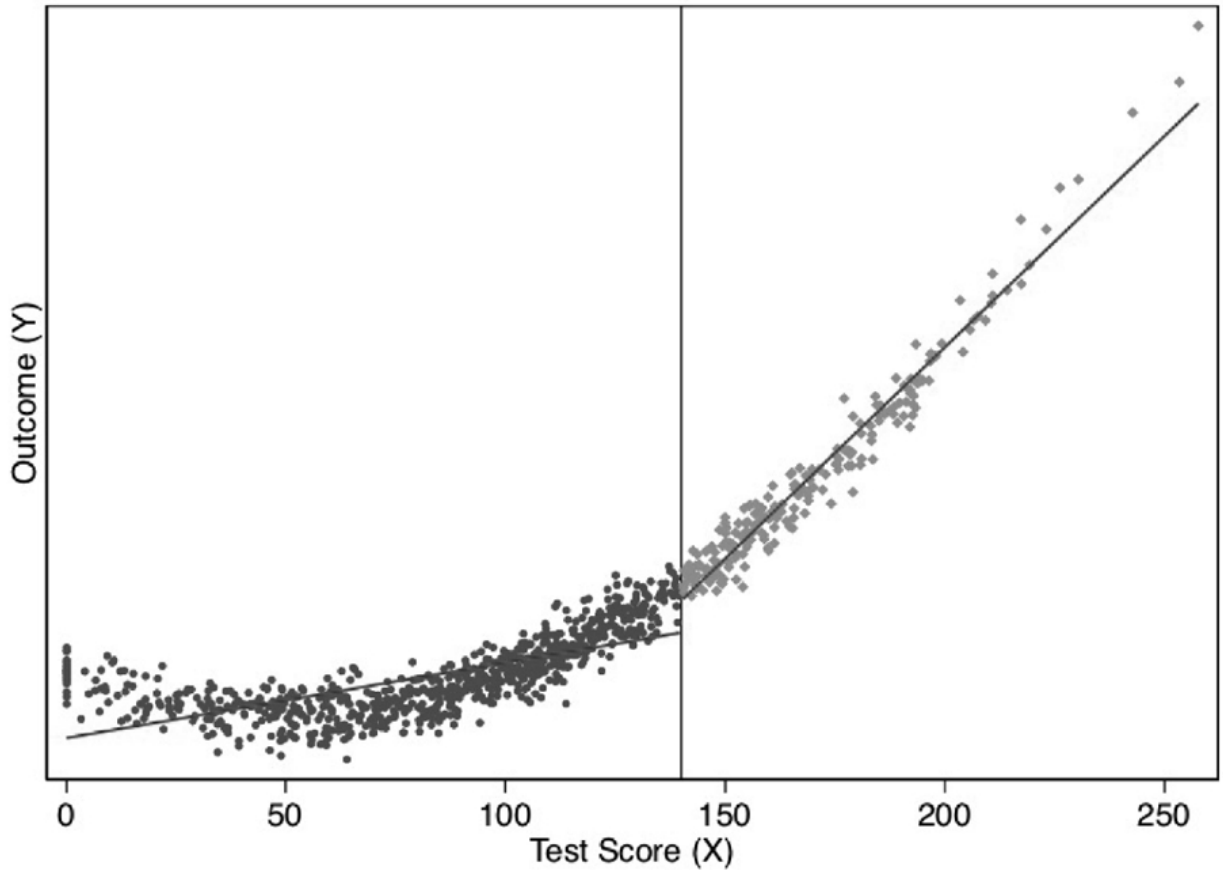


Figure 27. Simulated nonlinear data from Stata.

Table 40. Estimated effect of *D* on *Y* using OLS controlling for linear running variable.

Dependent variable	**Y**
Treatment (D)	6580.16***
	(305.88)

Higher-order polynomials can lead to overfitting and have been found to introduce bias [Gelman and Imbens, 2019]. 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 discuss later.

Though Gelman and Imbens [2019] warn us about higher-order polynomials, I'd like to use an example with *p*th-order polynomials, mainly because it's not uncommon to see this done today. I'd also

like you to know some of the history of this method and understand better what old papers were doing. 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:

$$E[Y_i^0 \mid X_i] = \alpha + \beta_{01}\tilde{X}_i + \cdots + \beta_{0p}\tilde{X}_i^p$$
$$E[Y_i^1 \mid X_i] = \alpha + \delta + \beta_{11}\tilde{X}_i + \cdots + \beta_{1p}\tilde{X}_i^p$$

where $\tilde{X}_i$ is the recentered 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:

$$E[Y \mid X] = E[Y^0 \mid X] + \Big(E[Y^1 \mid X] - E[Y^0 \mid X]\Big)D$$

Your regression model then is

$$Y_i = \alpha + \beta_{01}\tilde{X}_i + \cdots + \beta_{0p}\tilde{X}_i^p + \delta D_i + \beta_1^* D_i\tilde{X}_i + \cdots + \beta_p^* D_i\tilde{X}_i^p + \varepsilon_i$$

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 $\delta + \beta_1^* c + \cdots + \beta_p^* c^p$. Let's see this in action with another simulation.

STATA

rdd_simulate4.do

```
* Polynomial modeling
capture drop y
gen y = 10000 + 0*D - 100*x +x2 + rnormal(0, 1000)
reg y D##c.(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)")
```

Table 41. Estimated effect of *D* on *Y* using OLS controlling for linear and quadratic running variable.

Dependent variable	**Y**
Treatment (D)	-43.24
	(147.29)

R

rdd_simulate4.R

```
library(stargazer)

dat <- tibble(
  x = rnorm(1000, 100, 50)
) %>%
  mutate(
    x = case_when(x < 0 ~ 0, TRUE ~ x),
    D = case_when(x > 140 ~ 1, TRUE ~ 0),
    x2 = x*x,
    x3 = x*x*x,
    y3 = 10000 + 0 * D - 100 * x + x2 + rnorm(1000, 0, 1000)
  ) %>%
  filter(x < 280)

regression <- lm(y3 ~ D*., data = dat)

stargazer(regression, type = "text")

ggplot(aes(x, y3, colour = factor(D)), data = dat) +
  geom_point(alpha = 0.2) +
  geom_vline(xintercept = 140, colour = "grey", linetype = 2) +
  stat_smooth(method = "loess", se = F) +
  labs(x = "Test score (X)", y = "Potential Outcome (Y)")
```

Let's look at the output from this exercise in Figure 28 and Table 41. As you can see, once we model the data using a quadratic (the cubic ultimately was unnecessary), there is no estimated treatment effect at the cutoff. There is also no effect in our least squares regression. *Nonparametric kernels*. But, as we mentioned earlier, Gelman and Imbens [2019] have discouraged the use of higher-order polynomials when estimating local linear regressions. An 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 29). In this picture, the bias is caused by strong

trends in expected potential outcomes throughout the running variable.

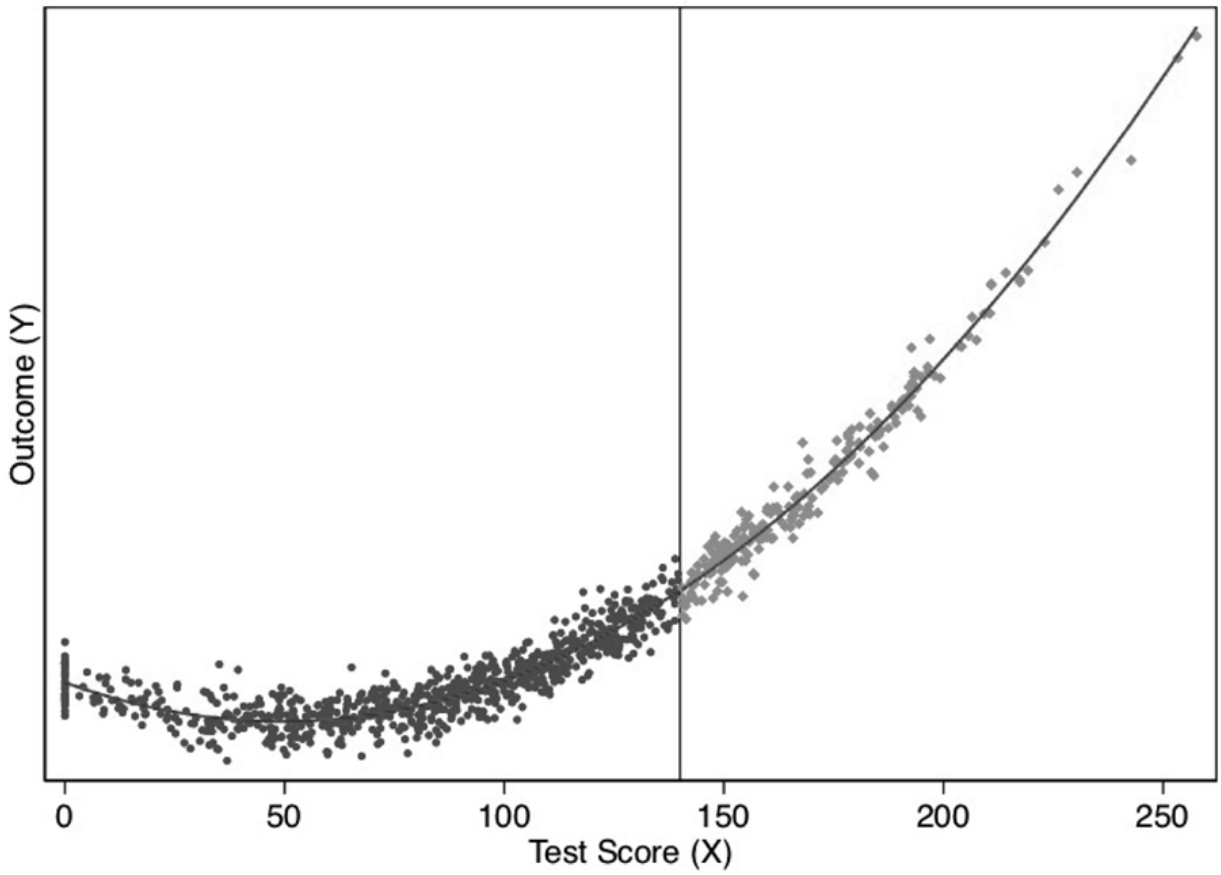


Figure 28. Simulated nonlinear data from Stata.

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-xsloping.

The standard solution to this problem is to run local linear nonparametric 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.[7] 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.

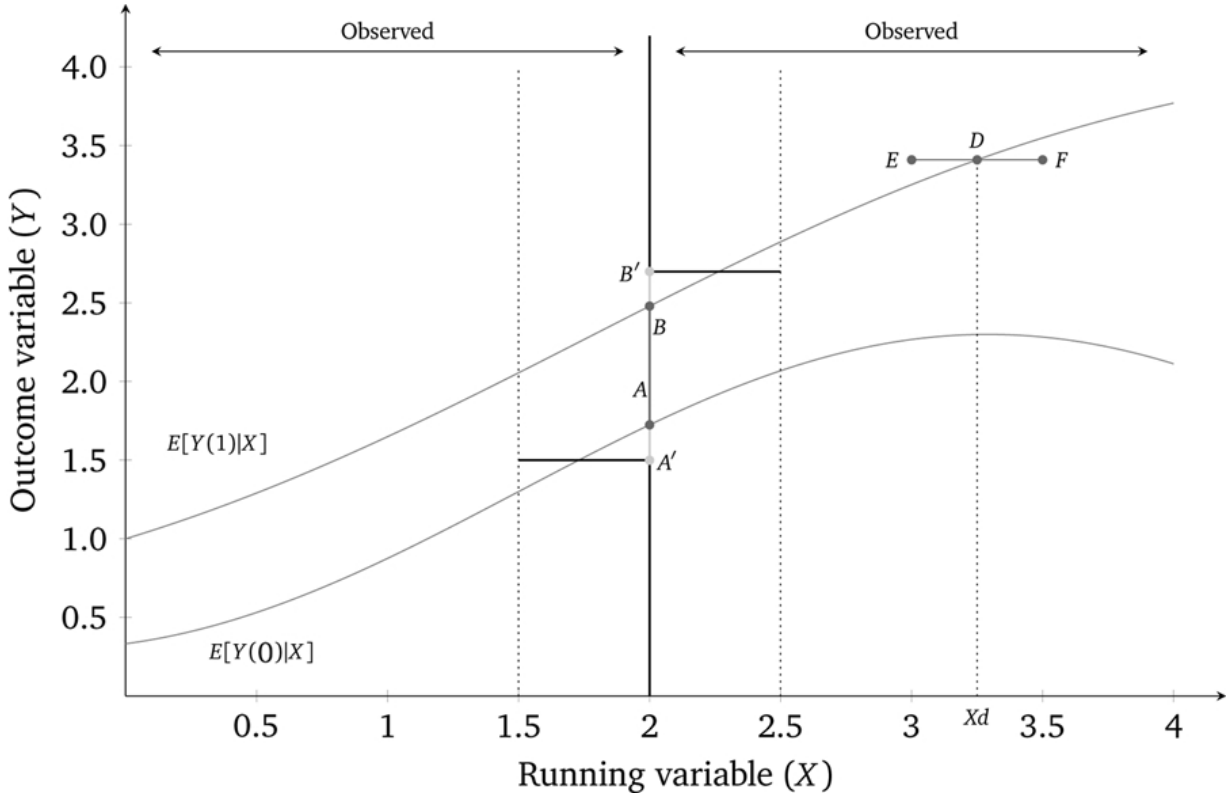


Figure 29. Boundary problem.

The model is some version of:

$$(\hat{a},\hat{b}) =_{a,b} \sum_{i=1}^{n} \left(y_i - a - b(x_i - c_0)\right)^2 K\left(\frac{x_i - c_o}{h}\right) 1(x_i > c_0) \qquad (6.4)$$

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 bandwidth. This method is sensitive to the choice of bandwidth, but more recent work allows the researcher to estimate *optimal* bandwidths [Calonico et al., 2014; Imbens and Kalyanaraman, 2011]. These may even allow for bandwidths to vary left and right of the cutoff.

Medicare and universal health care. 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? Universal insurance has become highly relevant because of the debates surrounding the Affordable Care Act, as well as several Democratic senators supporting Medicare for All. But it is also important for its sheer size. In 2014, Medicare was 14% of the federal budget at $505 billion.

Approximately 20% of non-elderly adults in the United States 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 policies, 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 is uninsured. Most have fee-for-service Medicare coverage. And that transition to Medicare occurs sharply at 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 $(k = 1, 2, \ldots, K)$ 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: $cov(u,C) = 0$. 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.

Suppose health insurance coverage can be summarized by two dummy variables: C^1_{ija} (any coverage) and C^2_{ija} (generous insurance). Card et al. [2008] estimate the following linear probability models:

$$C^1_{ija} = X_{ija}\beta^1_j + g^1_j(a) + D_a\pi_j{}^1 + v^1_{ija}$$
$$C^2_{ija} = X_{ija}\beta^2_j + g^2_j(a) + D_a\pi_j{}^2 + v^2_{ija}$$

where β^1_j and β^2_j are group-specific coefficients, $g^1_j(a)$ and $g^2_j(a)$ 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:

$$y_{ija} = X_{ija}\alpha + f_k(\alpha;\beta) + \sum_k C^k_{ija}\delta^k + u_{ija}$$

Combining the C_{ija} equations, and rewriting the reduced form model, we get:

$$y_{ija} = X_{ija}\left(\alpha_j + \beta^1_j\delta^1_j + \beta^2_j\delta^2_j\right)h_j(a) + D_a\pi^y_j + v^y_{ija}$$

where $h(a) = f_i(a) + \delta^1 g^1_i(a) + \delta^2 g^2_i(a)$ the reduced form age profile for group j, $\pi^y_j = \pi_j{}^1\delta^1 + \pi_j{}^2\delta^2$ and $v^y_{ija} = u_{ija} + v^1_{ija}\delta^1 + v^2_{ija}\delta^2$ is the error term. Assuming that the profiles $f_j(a)$, $g_j(a)$, and $g^2_j(a)$ 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 $(\pi_j{}^1$ and $\pi_j{}^2)$ 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:

$$\pi_j^y = \delta^0 + \delta^1 \pi_j^1 + \delta_j^2 \pi_j^2 + e_j$$

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 data sets—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 will 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 data set 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 forty 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 they turn 65 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 age 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 (age 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 42. The way that you read this table is each cell shows the *average treatment effect* for the 65-year-old 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 the conditional expectation functions for both potential outcomes is continuous at age=65. This means that both $E[Y^0 \mid a]$ and $E[Y^1 \mid a]$

are continuous through age of 65. If that assumption is plausible, then the average treatment effect at age 65 is identified as:

$$\lim_{65 \leftarrow a} E[y^1 \mid a] - \lim_{a \rightarrow 65} E[y^0 \mid a]$$

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 nonworkers 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 data set—the March CPS 1996–2004. And they ultimately find no evidence for discontinuities in employment at age 65 (Figure 30).

Table 42. Insurance characteristics just before age 65 and estimated discontinuities at age 65.

	On Medicare	Any insurance	Private coverage	2+ forms coverage	Managed care
Overall sample	59.7	9.5	−2.9	44.1	−28.4
	(4.1)	(0.6)	(1.1)	(2.8)	(2.1)
White non-Hispanic					
Less than high school	58.5	13.0	−6.2	44.5	−25.0
	(4.6)	(2.7)	(3.3)	(4.0)	(4.5)
High school graduate	64.7	7.6	−1.9	51.8	−30.3
	(5.0)	(0.7)	(1.6)	(3.8)	(2.6)
Some college	68.4	4.4	−2.3	55.1	−40.1
	(4.7)	(0.5)	(1.8)	(4.0)	(2.6)
Minority					
High school dropout	44.5	21.5	−1.2	19.4	−8.3
	(3.1)	(2.1)	(2.5)	(1.9)	(3.1)
High school graduate	44.6	8.9	−5.8	23.4	−15.4
	(4.7)	(2.8)	(5.1)	(4.8)	(3.5)
Some college	52.1	5.8	−5.4	38.4	−22.3
	(4.9)	(2.0)	(4.3)	(3.8)	(7.2)
Classified by ethnicity only					
White non-Hispanic	65.2	7.3	−2.8	51.9	−33.6
	(4.6)	(0.5)	(1.4)	(3.5)	(2.3)
Black non-Hispanic	48.5	11.9	−4.2	27.8	−13.5
	(3.6)	(2.0)	(2.8)	(3.7)	(3.7)
Hispanic	44.4	17.3	−2.0	21.7	−12.1
	(3.7)	(3.0)	(1.7)	(2.1)	(3.7)

Note: Entries in each cell are estimated regression discontinuities at age 65 from quadratics in age interacted with a dummy for 65 and older. Other controls such as gender, race, education, region, and sample year are also included. Data is from the pooled 1999–2003 NHIS.

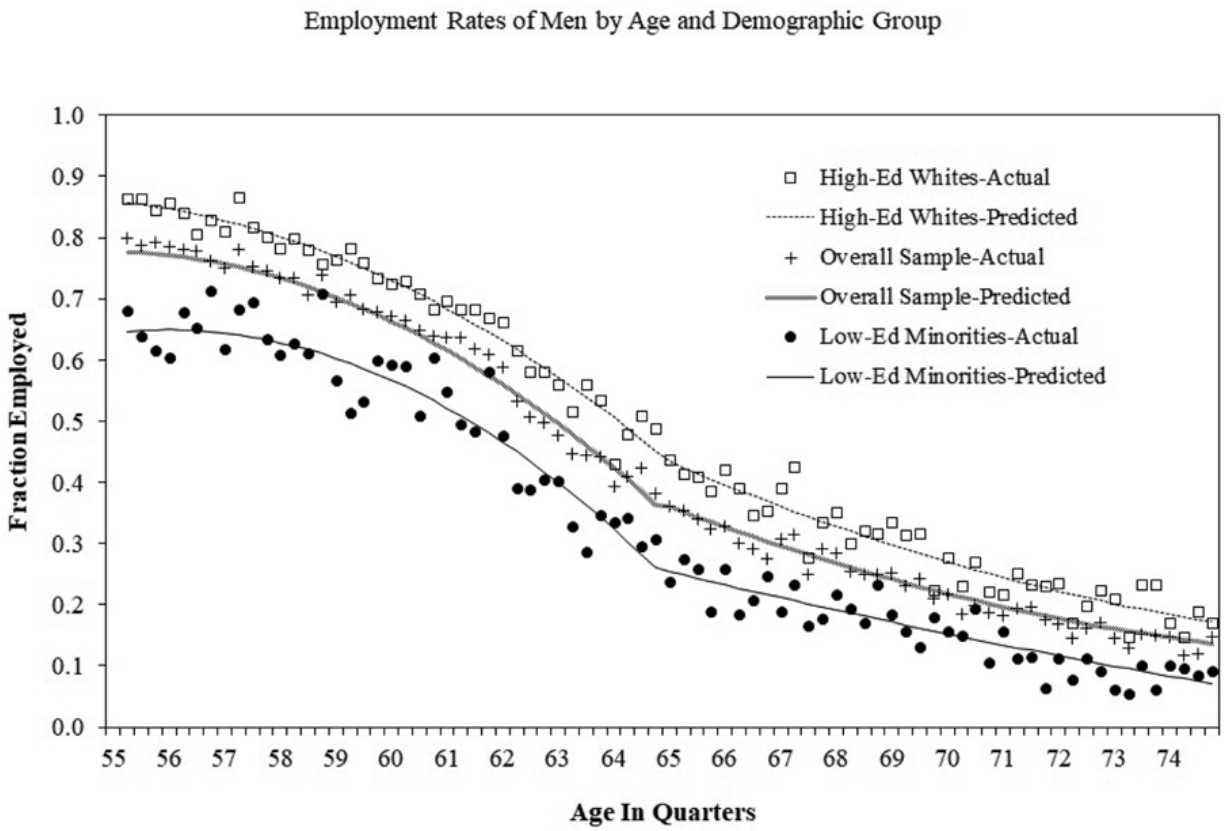


Figure 30. Investigating the CPS for discontinuities at age 65 [Card et al., 2008].

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?"

"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 presented in Table 43. Each cell measures the average treatment effect for the complier population at the discontinuity. Standard errors are in parentheses. There are 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.

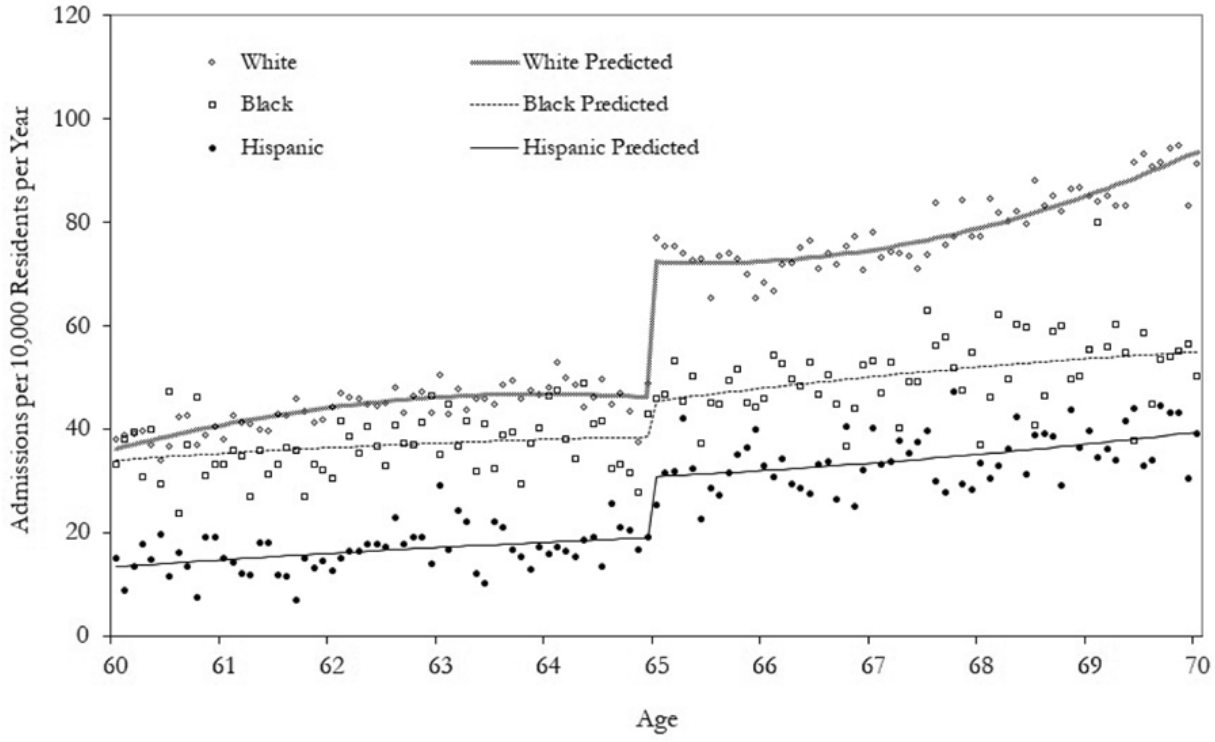


Figure 31. Changes in hospitalizations [Card et al., 2008].

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 31 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 health-care 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 found slight decreases in mortality rates (see Table 44). *Inference*. As we've mentioned, it's standard practice in the RDD to estimate causal effects using local polynomial regressions. In its simplest form, this amounts to nothing

more complicated than fitting a linear specification separately on each side of the cutoff using a least squares regression. But when this is done, you are using only the observations within some pre-specified window (hence “local”). As the true conditional expectation function is probably not linear at this window, the resulting estimator likely suffers from specification bias. But if you can get the window narrow enough, then the bias of the estimator is probably small relative to its standard deviation.

Table 43. Measures of access to care just before 65 and estimated discontinuities at age 65.

	Delayed last year	Did not get care last year	Saw doctor last year	Hospital stay last year
Overall sample	−1.8	−1.3	1.3	1.2
	(0.4)	(0.3)	(0.7)	(0.4)
White non-Hispanic				
Less than high school	−1.5	−0.2	3.1	1.6
	(1.1)	(1.0)	(1.3)	(1.3)
High school graduate	0.3	−1.3	−0.4	0.3
	(2.8)	(2.8)	(1.5)	(0.7)
Some college	−1.5	−1.4	0.0	2.1
	(0.4)	(0.3)	(1.3)	(0.7)
Minority				
High school dropout	−5.3	−4.2	5.0	0.0
	(1.0)	(0.9)	(2.2)	(1.4)
High school graduate	−3.8	1.5	1.9	1.8
	(3.2)	(3.7)	(2.7)	(1.4)
Some college	−0.6	−0.2	3.7	0.7
	(1.1)	(0.8)	(3.9)	(2.0)
Classified by ethnicity only				
White non-Hispanic	−1.6	−1.2	0.6	1.3
	(0.4)	(0.3)	(0.8)	(0.5)
Black non-Hispanic	−1.9	−0.3	3.6	0.5
	(1.1)	(1.1)	(1.9)	(1.1)
Hispanic	−4.9	−3.8	8.2	11.8
	(0.8)	(0.7)	(0.8)	(1.6)

Note: Entries in each cell are estimated regression discontinuities at age 65 from quadratics in age interacted with a dummy for 65 and older.Other controls such as gender, race, education, region and sample year are also included. First two columns are from 1997–2003 NHIS and last two columns are from 1992–2003 NHIS.

Table 44. Regression discontinuity estimates of changes in mortality rates.

	Death rate in					
	7 days	14 days	28 days	90 days	180 days	365 days
Quadratic	−1.1	−1.0	−1.1	−1.2	−1.0	
no controls	(0.2)	(0.2)	(0.3)	(0.3)	(0.4)	(0.4)
Quadratic	−1.0	−0.8	−0.9	−0.9	−0.8	−0.7
plus controls	0.2)	(0.2)	(0.3)	(0.3)	(0.3)	(0.4)
Cubic plus	−0.7	−0.7	−0.6	−0.9	−0.9	−0.4
controls	(0.3)	(0.2)	(0.4)	(0.4)	(0.5)	(0.5)
Local OLS with ad	−0.8	−0.8	−0.8	−0.9	−1.1	−0.8
hoc bandwidths	(0.2)	(0.2)	(0.2)	(0.2)	(0.3)	(0.3)

Note: Dependent variable for death within interval shown in the column heading. Regression estimates at the discontinuity of age 65 for flexible regression models. Standard errors in parentheses.

But what if the window cannot be narrowed enough? This can happen if the running variable only takes on a few values, or if the gap between values closest to the cutoff are large. The result could be you simply do not have enough observations close to the cutoff for the local polynomial regression. This also can lead to the heteroskedasticity-robust confidence intervals to undercover the average causal effect because it is not centered. And here's the really bad news—this probably is happening *a lot* in practice.

In a widely cited and very influential study, Lee and Card [2008] suggested that researchers should cluster their standard errors by the running variable. This advice has since become common practice in the empirical literature. Lee and Lemieux [2010], in a survey article on proper RDD methodology, recommend this practice, just to name one example. But in a recent study, Kolesár and Rothe [2018] provide extensive theoretical and simulation-based evidence that clustering on the running variable is perhaps one of the *worst* approaches you could take. In fact, clustering on the running variable can actually be substantially worse than heteroskedastic-robust standard errors.

As an alternative to clustering and robust standard errors, the authors propose two alternative confidence intervals that have guaranteed coverage properties under various restrictions on the conditional expectation function. Both confidence intervals are "honest," which means they achieve correct coverage uniformly over all conditional expectation functions in large samples. These confidence intervals are currently unavailable in Stata as of the time of this writing, but they can be implemented in R with the RDHonest package.[8] R users are encouraged to use these confidence intervals. Stata users are encouraged to switch (grudgingly) to R so as to use these confidence intervals. Barring that, Stata users should use the heteroskedastic robust standard errors. But whatever you do, don't cluster on the running variable, as that is nearly an unambiguously bad idea.

A separate approach may be to use randomization inference. As we noted, Hahn et al. [2001] emphasized that the conditional expected potential outcomes must be continuous across the cutoff for a regression discontinuity design to identify the local average treatment effect. But Cattaneo et al. [2015] suggest an alternative assumption which has implications for inference. They ask us to consider that perhaps around the cutoff, in a short enough window, the treatment was assigned to units randomly. It was effectively a coin flip which side of the cutoff someone would be for a small enough window around the cutoff. Assuming there exists a neighborhood around the cutoff where this randomization-type condition holds, then this assumption may be viewed as an approximation of a randomized experiment around the cutoff. Assuming this is plausible, we can proceed as if only those observations closest to the discontinuity were randomly assigned, which leads naturally to randomization inference as a methodology for conducting exact or approximate *p*-values.

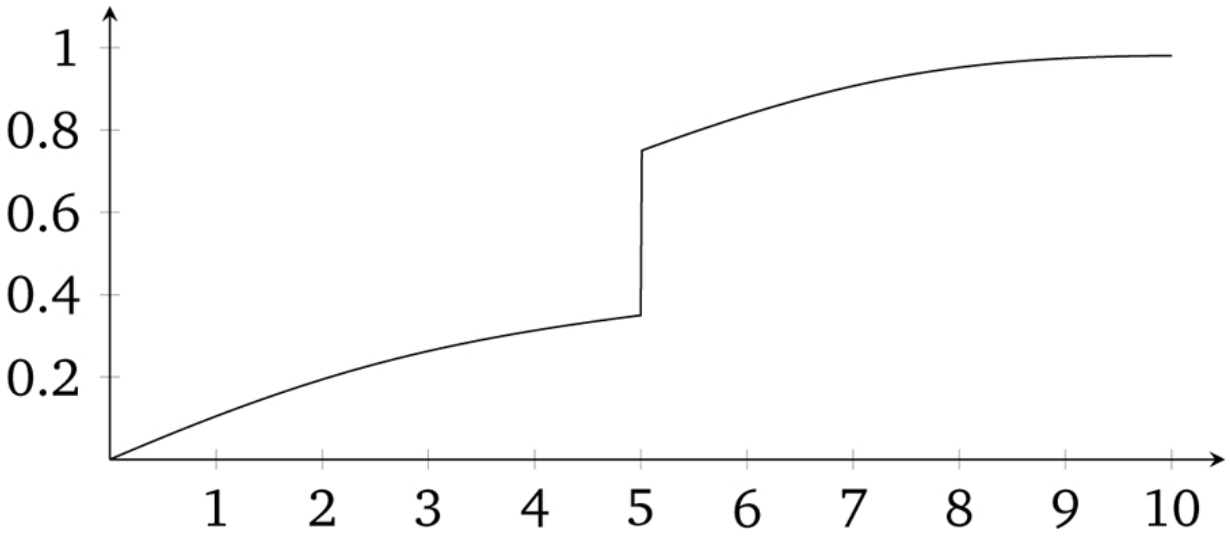


Figure 32. Vertical axis is the probability of treatment for each value of the running variable.

The Fuzzy RD Design. 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 earlier paper by Hoekstra [2009] had this feature, as did Angrist and Lavy [1999]. The formal definition of a probabilistic treatment assignment is

$$\lim_{X_i \to c_0} \Pr(D_i = 1 \mid X_i = c_0) \neq \lim_{c_0 \leftarrow X_i} \Pr(D_i = 1 \mid X_i = c_0) \qquad (6.5)$$

In other words, the conditional probability is discontinuous as X approaches c_0 in the limit. A visualization of this is presented from Imbens and Lemieux [2008] in Figure 32.

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 33.

Estimating some average treatment effect under a fuzzy RDD is very similar to how we estimate a local average treatment effect with instrumental variables. I will cover instrumental variables in more detail later in the book, but for now let me tell you about estimation under fuzzy designs using IV. One can estimate several ways. One simple way is a type of Wald estimator, where you estimate some causal effect as 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.

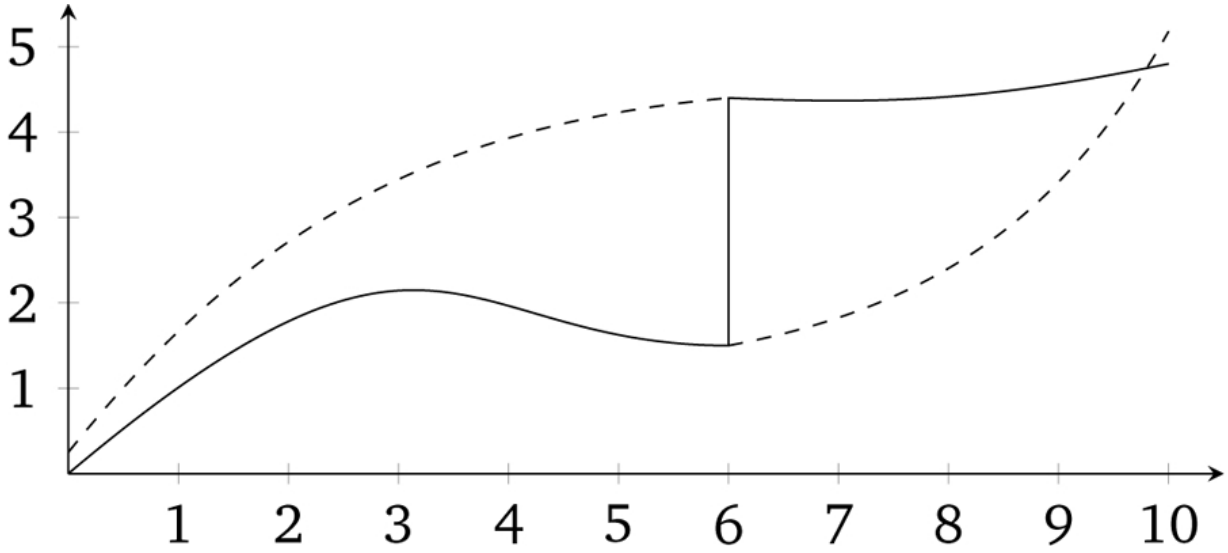


Figure 33. Potential and observed outcomes under a fuzzy design.

$$\delta_{\text{Fuzzy RDD}} = \frac{\lim_{X \to c_0} E[Y \mid X = c_0] - \lim_{X_0 \leftarrow X} E[Y \mid X = c_0]}{\lim_{X \to c_0} E[D \mid X = c_0] - \lim_{X_0 \leftarrow X} E[D \mid X = c_0]} \tag{6.6}$$

The assumptions for identification here are the same as with any instrumental variables design: all the caveats about exclusion restrictions, monotonicity, SUTVA, and the strength of the first stage.[9]

But one can also estimate the effect using a two-stage least squares model or similar appropriate model such as limited-information maximum likelihood. Recall that there are now two

events: the first event is when the running variable exceeds the cutoff, and the second event is when a unit is placed in the treatment. Let Z_i be an indicator for when X exceeds c_0. One can use both Z_i and the interaction terms as instruments for the treatment D_i. If one uses only Z_i as an instrumental variable, then it is a "just identified" model, which usually has good finite sample properties.

Let's look at a few of the regressions that are involved in this instrumental variables approach. There are three possible regressions: the first stage, the reduced form, and the second stage. Let's look at them in order. In the case just identified (meaning only one instrument for one endogenous variable), the first stage would be:

$$D_i = \gamma_0 + \gamma_1 X_i + \gamma_2 X_i^2 + \cdots + \gamma_p X_i^p + \pi Z_i + \zeta_{1i}$$

where π is the causal effect of Z_i on the conditional probability of treatment. The fitted values from this regression would then be used in a second stage. We can also use both Z_i and the interaction terms as instruments for D_i. If we used Z_i and all its interactions, the estimated first stage would be:

$$D_i = \gamma_{00} + \gamma_{01}\tilde{X}_i + \gamma_{02}\tilde{X}_i^2 + \cdots + \gamma_{0p}\tilde{X}_i^p + \pi Z_i + \gamma_1^*\tilde{X}_i Z_i + \gamma_2^*\tilde{X}_i Z_i + \cdots + \gamma_p^* Z_i + \zeta_{1i}$$

We would also construct analogous first stages for $\tilde{X}_i D_i, \ldots, \tilde{X}_i^p D_i$.

If we wanted to forgo estimating the full IV model, we might estimate the reduced form only. You'd be surprised how many applied people prefer to simply report the reduced form and not the fully specified instrumental variables model. If you read Hoekstra [2009], for instance, he favored presenting the reduced form—that second figure, in fact, was a picture of the reduced form. The reduced form would regress the outcome Y onto the instrument and the running variable. The form of this fuzzy RDD reduced form is:

$$Y_i = \mu + \kappa_1 X_i + \kappa_2 X_i^2 + \cdots + \kappa_p X_i^p + \delta\pi Z_i + \zeta_{2i}$$

As in the sharp RDD case, one can allow the smooth function to be different on both sides of the discontinuity by interacting Z_i with the running variable. The reduced form for this regression is:

$$Y_i = \mu + \kappa_{01} X_i \tilde{X}_i + \kappa_{02} X_i \tilde{X}_i^2 + \cdots + \kappa_{0p} X_i \tilde{X}_i^p + \delta\pi Z_i + \kappa_{01} X_i^* \tilde{X}_i Z_i + \kappa_{02} X_i^* \tilde{X}_i Z_i + \cdots + \kappa_{0p} X_i^* Z_i + \zeta_{1i}$$

But let's say you wanted to present the estimated effect of the treatment on some outcome. That requires estimating a first stage, using fitted values from that regression, and then estimating a second stage on those fitted values. This, and only this, will identify the causal effect of the treatment on the outcome of interest. The reduced form only estimates the causal effect of the instrument on the outcome. The second-stage model with interaction terms would be the same as before:

$$Y_i = \alpha + \beta_{01} \tilde{x}_i + \beta_{02} \tilde{x}_i^2 + \cdots + \beta_{0p} \tilde{x}_i^p + \delta \widehat{D}_i + \beta_1^* \widehat{D}_i \tilde{x}_i + \beta_2^* \widehat{D}_i \tilde{x}_i^2 + \cdots + \beta_p^* \widehat{D}_i \tilde{x}_i^p + \eta_i$$

Where $\tilde{x}$ are now not only normalized with respect to c_0 but are also fitted values obtained from the first-stage regressions.

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 requirement 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 the only thing that causes the

outcome to change abruptly at c_0 is the treatment. But, this can be violated in practice if any of the following is true:

1. The assignment rule is known in advance.
2. Agents are interested in adjusting.
3. Agents have time to adjust.
4. The cutoff is endogenous to factors that independently cause potential outcomes to shift.
5. There is nonrandom heaping along the running variable.

Examples include retaking an exam, self-reporting income, and so on. 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 years old and faces more severe penalties for crime. This age threshold triggers the treatment (i.e., higher penalties for crime), but is also correlated with variables that affect the outcomes, such as graduating from high school and voting rights. Let's tackle these problems separately.

McCrary's density test. Because of these challenges to identification, a lot of work by econometricians and applied microeconomists has gone toward trying to figure out solutions to these problems. The most influential is a density test by Justin McCrary, now called the McCrary density test [2008]. The McCrary density test is used to check whether units are sorting on the running variable. Imagine that there are two rooms with patients in line for some life-saving treatment. Patients in room A will receive the life-saving treatment, and patients in room B will *knowingly* receive nothing. What would you do if you were in room B? Like me, you'd probably stand up, open the door, and walk across the hall to room A. There are natural incentives for the people in room B to get into room A, and the only thing that would keep people in room B from sorting into room A is if doing so were impossible.

But, let's imagine that the people in room B had successfully sorted themselves into room A. What would that look like to an

outsider? If they were successful, then room A would have more patients than room B. In fact, in the extreme, room A is crowded and room B is empty. This is the heart of the McCrary density test, and when we see such things at the cutoff, we have some suggestive evidence that people are sorting on the running variable. This is sometimes called manipulation.

Remember earlier when I said we should think of continuity as the null because nature doesn't make jumps? If you see a turtle on a fencepost, it probably didn't get there itself. Well, the same goes for the density. If the null is a continuous density through the cutoff, then bunching in the density at the cutoff is a sign that someone is moving over to the cutoff—probably to take advantage of the rewards that await there. Sorting on the sorting variable is a testable prediction under the null of a continuous density. Assuming a continuous distribution of units, sorting on the running variable means that units are moving just on the other side of the cutoff. Formally, if we assume a desirable treatment D and an assignment rule $X \geq c_0$, then we expect individuals will sort into D by choosing X such that $X \geq c_0$ —so long as they're able. If they do, then it could imply selection bias insofar as their sorting is a function of potential outcomes.

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.[10] I've always liked this test because it's a really simple statistical test based on a theory that human beings are optimizing under constraints. And if they are optimizing, that makes for testable predictions—like a discontinuous jump in the density at the cutoff. Statistics built on behavioral theory can take us further.

To implement the McCrary density test, 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. I recommend the package rddensity,[11] which you can install for R as well.[12] These packages are based on Cattaneo et al. [2019], which is based on local polynomial regressions that have less bias in the border regions.

This is a high-powered test. You need a lot of observations at c_0 to distinguish a discontinuity in the density from noise. Let me illustrate in Figure 34 with a picture from McCrary [2008] that shows a situation with and without manipulation.

Covariate balance and other placebos. It has become common in this literature to provide evidence for the credibility of the underlying identifying assumptions, at least to some degree. While the assumptions cannot be directly tested, indirect evidence may be persuasive. I've 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 reasonably chosen covariates around the cutoff. As these are pretreatment characteristics, they should be invariant to change in treatment assignment. An example of this is from Lee et al. [2004], who evaluated the impact of Democratic vote share just at 50%, on various demographic factors (Figure 35).

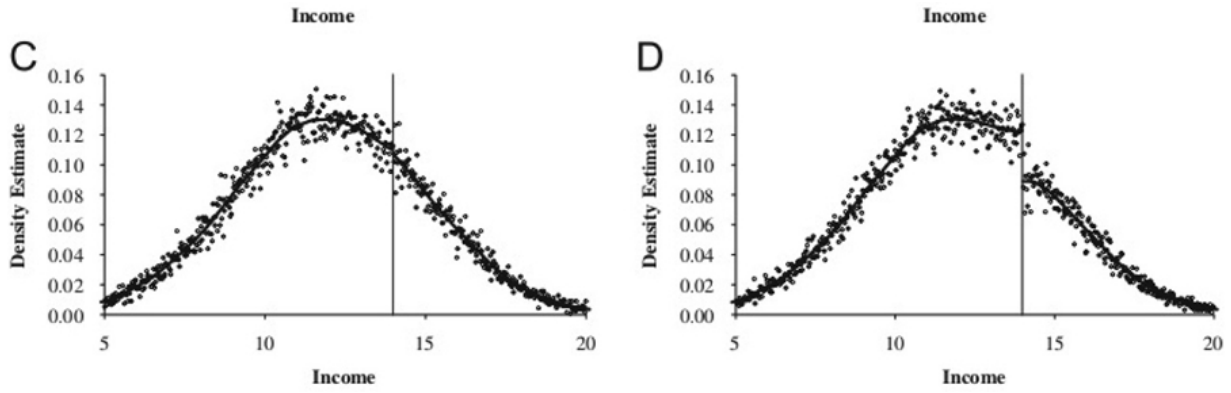


Figure 34. A picture with and without a discontinuity in the density. Reprinted from *Journal of Econometrics*, 142, J. McCrary, "Manipulation of the Running Variable in the Regression Discontinuity Design: A Design Test," 698–714. © 2008, with permission from Elsevier.

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 looking at one side of the discontinuity, taking the median value of the running variable in that section, and pretending 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.

Nonrandom heaping on the running variable. Almond et al. [2010] is a fascinating study. The authors are interested in estimating the causal effect of medical expenditures on health outcomes, in part because many medical technologies, while effective, may not justify the costs associated with their use. Determining their effectiveness is challenging given that medical resources are, we hope, optimally assigned to patients based on patient potential outcomes. To put it a different way, if the physician perceives that an intervention will have the best outcome, then that is likely a treatment that will be assigned to the patient. This violates independence, and more than likely, if the endogeneity of the treatment is deep enough, controlling for selection directly will be tough, if not impossible. As we saw with our

earlier example of the perfect doctor, such nonrandom assignment of interventions can lead to confusing correlations. Counterintuitive correlations may be nothing more than selection bias.

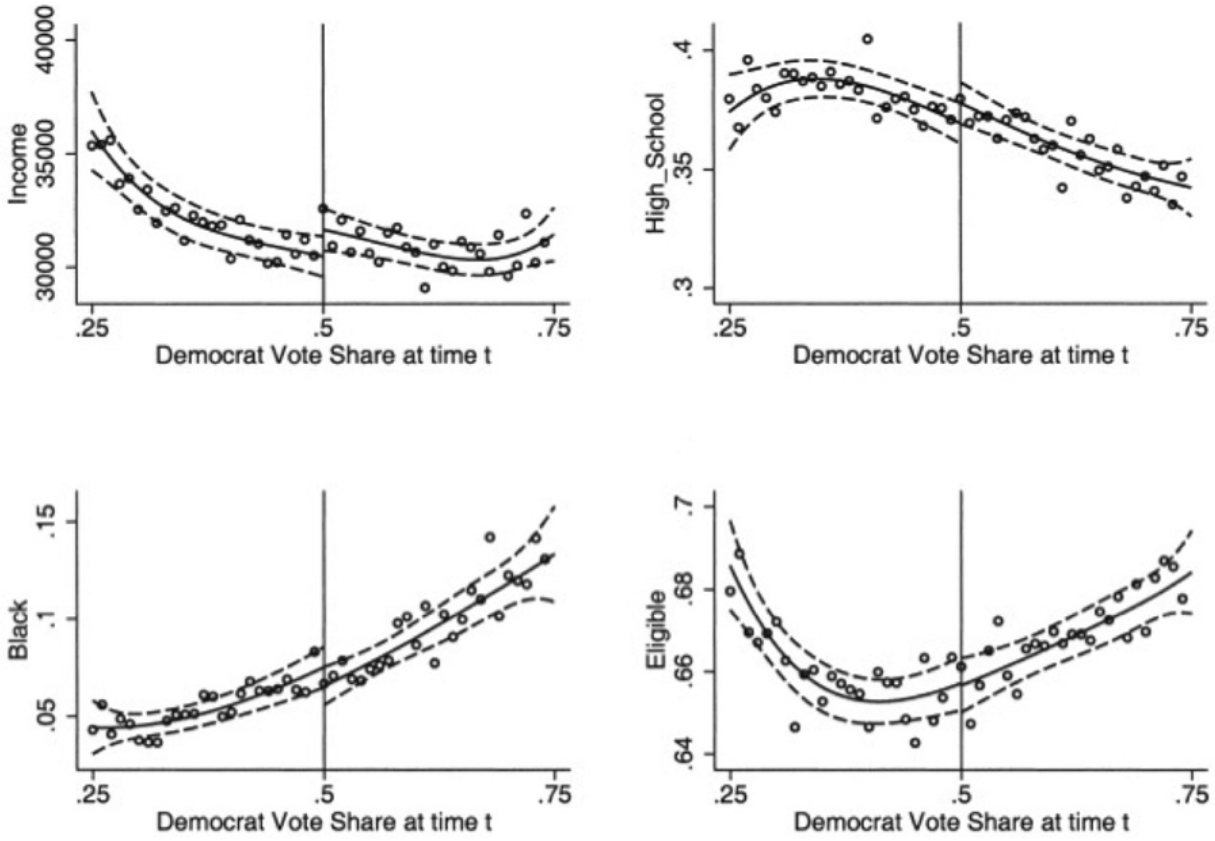


Figure 35. Panels refer to (top left to bottom right) district characteristics: real income, percentage high school degree, percentage black, and percentage eligible to vote. Circles represent the average characteristic within intervals of 0.01 in Democratic vote share. The continuous line represents the predicted values from a fourth-order polynomial in vote share fitted separately for points above and below the 50% threshold. The dotted line represents the 95% confidence interval. Reprinted from Lee, D.S., Moretti, E., and Butler, M. J. (2004). "Do Voters Affect or Elect Policies: Evidence from the U.S. House." *Quarterly Journal of Economics*, 119(3):807–859. Permission from Oxford University Press.

But Almond et al. [2010] had an ingenious insight—in the United States, it is typically the case that babies with a very low birth weight receive heightened medical attention. This categorization is called the "very low birth weight" range, and such low birth weight is quite

dangerous for the child. Using administrative hospital records linked to mortality data, the authors find that the 1-year infant mortality decreases by around 1 percentage point when the child's birth weight is just below the 1,500-gram threshold compared to those born just above. Given the mean 1-year mortality of 5.5%, this estimate is sizable, suggesting that the medical interventions triggered by the very-low-birth-weight classification have benefits that far exceed their costs.

Barreca et al. [2011] and Barreca et al. [2016] highlight some of econometric issues related to what they call "heaping" on the running variable. Heaping is when there is an excess number of units at certain points along the running variable. In this case, it appeared to be at regular 100-gram intervals and was likely caused by a tendency for hospitals to round to the nearest integer. A visualization of this problem can be seen in the original Almond et al. [2010], which I reproduce here in Figure 36. The long black lines appearing regularly across the birth-weight distribution are excess mass of children born at those numbers. This sort of event is unlikely to occur naturally in nature, and it is almost certainly caused by either sorting or rounding. It could be due to less sophisticated scales or, more troubling, to staff rounding a child's birth weight to 1,500 grams in order to make the child eligible for increased medical attention.

Almond et al. [2010] attempt to study this more carefully using the conventional McCrary density test and find no clear, statistically significant evidence for sorting on the running variable at the 1,500-gram cutoff. Satisfied, they conduct their main analysis, in which they find a causal effect of around a 1-percentage-point reduction in 1-year mortality.

The focus of Barreca et al. [2011] and Barreca et al. [2016] is very much on the heaping phenomenon shown in Figure 36. Part of the strength of their work, though, is their illustration of some of the shortcomings of a conventional McCrary density test. In this case, the data heap at 1,500 grams appears to be babies whose mortality rates are unusually high. These children are outliers compared to units to *both* the immediate left and the immediate right. It is important to note that such events would not occur naturally; there is

no reason to believe that nature would produce heaps of children born with outlier health defects every 100 grams. The authors comment on what might be going on:

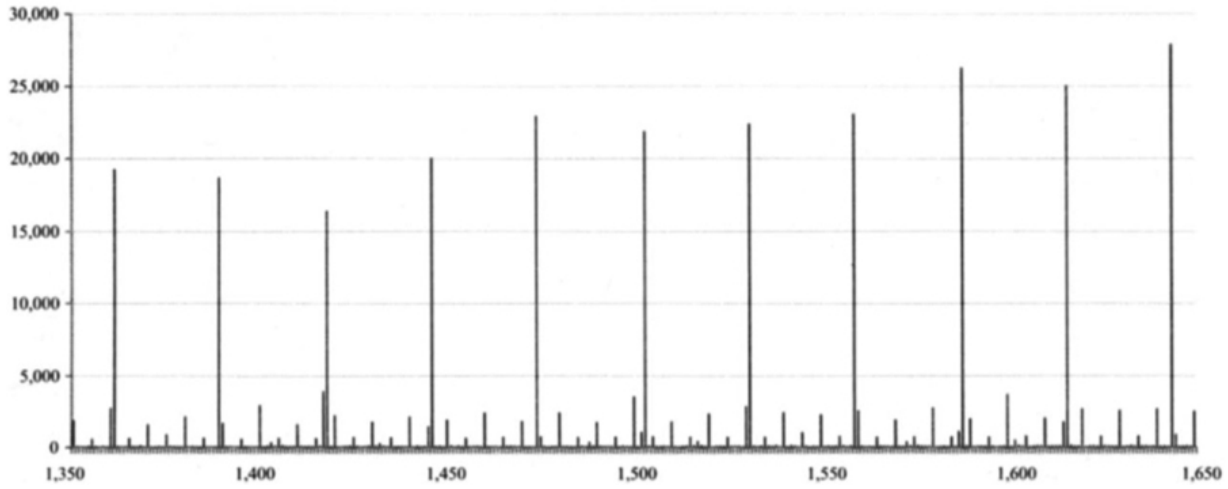


Figure 36. Distribution of births by gram. Reprinted from Almond, D., Doyle, J. J., Kowalski, A., and Williams, H. (2010). "Estimating Returns to Medical Care: Evidence from at-risk Newborns." *The Quarterly Journal of Economics*, 125(2):591–634. Permission from Oxford University Press.

> This [heaping at 1,500 grams] may be a signal that poor-quality hospitals have relatively high propensities to round birth weights but is also consistent with manipulation of recorded birth weights by doctors, nurses, or parents to obtain favorable treatment for their children. Barreca et al. [2011] show that this nonrandom heaping leads one to conclude that it is "good" to be strictly less than any 100-g cutoff between 1,000 and 3,000 grams.

Since estimation in an RDD compares means as we approach the threshold from either side, the estimates should not be sensitive to the observations at the thresholds itself. Their solution is a so-called "donut hole" RDD, wherein they remove units in the vicinity of 1,500 grams and reestimate the model. Insofar as units are dropped, the parameter we are estimating at the cutoff has become an even more unusual type of local average treatment effect that may be even less informative about the average treatment effects that policymakers are desperate to know. But the strength of this rule is that it allows for the possibility that units at the heap differ markedly due to selection bias from those in the surrounding area. Dropping these

units reduces the sample size by around 2% but has very large effects on 1-year mortality, which is approximately 50% lower than what was found by Almond et al. [2010].

These companion papers help us better understand some of the ways in which selection bias can creep into the RDD. Heaping is not the end of the world, which is good news for researchers facing such a problem. The donut hole RDD can be used to circumvent some of the problems. But ultimately this solution involves dropping observations, and insofar as your sample size is small relative to the number of heaping units, the donut hole approach could be infeasible. It also changes the parameter of interest to be estimated in ways that may be difficult to understand or explain. Caution with nonrandom heaping along the running variable is probably a good thing.

Replicating a Popular Design: The Close Election

Within RDD, there is a particular kind of design that has become quite popular, the close-election design. Essentially, this design exploits a feature of American democracies wherein winners in political races are declared when a candidate gets the minimum needed share of votes. Insofar as very close races represent exogenous assignments of a party's victory, which I'll discuss below, then we can use these close elections to identify the causal effect of the winner on a variety of outcomes. We may also be able to test political economy theories that are otherwise nearly impossible to evaluate.

The following section has two goals. First, to discuss in detail the close election design using the classic Lee et al. [2004]. Second, to show how to implement the close-election design by replicating several parts of Lee et al. [2004].

Do Politicians or Voters Pick Policies? The big question motivating Lee et al. (2004) has to do with whether and in which way voters affect policy. There are two fundamentally different views of the role

of elections in a representative democracy: convergence theory and divergence theory.

The convergence theory states that 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, 807]

Divergence theory is a slightly more commonsense view of political actors. When partisan politicians cannot credibly commit to certain policies, then convergence is undermined and the result can be full policy "divergence." Divergence is when the winning candidate, after taking office, simply pursues her most-preferred policy. In this extreme case, voters are unable to compel candidates to reach any kind of policy compromise, and this is expressed as two opposing candidates choosing very different policies under different counterfactual victory scenarios.

Lee et al. [2004] 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's 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 his or her 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 $x>y_e$, 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 equaled 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.[13]
   - 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.

The potential roll-call voting record outcomes of the candidate 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:

$$RC_t = \alpha_0 + \pi_0 P_t^* + \pi_1 D_t + \varepsilon_t$$

$$RC_{t+1} = \beta_0 + \pi_0 P_{t+1}^* + \pi_1 D_{t+1} + \varepsilon_{t+1}$$

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:

$$\underbrace{E[RC_{t+1} \mid D_t = 1] - E[RC_{t+1} \mid D_t = 0]}_{\text{Observable}} = \pi_0[P^{*D}_{t+1} - P^{*R}_{t+1}] + \underbrace{\pi_1[P^{D}_{t+1} - P^{R}_{t+1}]}_{\text{Observable}} = \underbrace{\gamma}_{\text{Total effect of initial win on future roll call votes}} \tag{6.7}$$

$$\underbrace{E[RC_t \mid D_t = 1] - E[RC_t \mid D_t = 0]}_{\text{Observable}} = \pi_1 \tag{6.8}$$

$$\underbrace{E[D_{t+1} \mid D_t = 1] - E[D_{t+1} \mid D_t = 0]}_{\text{Observable}} = P^{D}_{t+1} - P^{R}_{t+1} \tag{6.9}$$

The "elect" component is $\pi_1[P^{D}_{t+1} - P^{R}_{t+1}]$ and is 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 $[P^{D}_{t+1} - P^{R}_{t+1}]$. 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 "effect" 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.

Replication exercise. There are two main data sets 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 twenty-five 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 vote share. 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. If convergence is true, then Republicans and Democrats who just barely won should vote almost identically, whereas if divergence is true, they should vote differently at the margins of a close race. This "at the margins of a close race" is crucial because the idea is that it is at the margins of a close race that the distribution of voter preferences is the same. And if voter preferences are the same, but policies diverge at the cutoff, then it suggests politicians and not voters are driving policy making.

Table 45. Original results based on ADA scores—close elections sample.

Dependent variable	ADA_{t+1}	ADA_t	DEM_{t+1}
Estimated gap	21.2	47.6	0.48
	(1.9)	(1.3)	(0.02)

Note: Standard errors 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% and 52%. 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% and 52% and observations for which the Democrat vote share at time *t* is strictly between 48% and 50%. 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.

The exogenous shock comes from the discontinuity in the running variable. At a vote share of just above 0.5, the Democratic candidate wins. They argue that just around that cutoff, random chance determined the Democratic win—hence the random assignment of D_t [Cattaneo et al., 2015]. Table 45 is a reproduction of Cattaneo et al.'s main results. The effect of a Democratic victory increases liberal

voting by 21 points in the next period, 48 points in the current period, and the probability of reelection by 48%. The authors find evidence for both divergence and incumbency advantage using this design. Let's dig into the data ourselves now and see if we can find where the authors are getting these results. We will examine the results around Table 45 by playing around with the data and different specifications.

STATA

lmb_1.do

```
use https://github.com/scunning1975/mixtape/raw/master/lmb-data.dta, clear

* Replicating Table 1 of Lee, Moretti and Butler (2004)
reg score lagdemocrat    if lagdemvoteshare>.48 & lagdemvoteshare<.52,
    ↪ cluster(id)
reg score democrat       if lagdemvoteshare>.48 & lagdemvoteshare<.52,
    ↪ cluster(id)
reg democrat lagdemocrat if lagdemvoteshare>.48 & lagdemvoteshare<.52,
    ↪ cluster(id)
```

R

lmb_1.R

```
library(tidyverse)
library(haven)
library(estimatr)

read_data <- function(df)
{
 full_path <- paste("https://raw.github.com/scunning1975/mixtape/master/",
                    df, sep = "")
 df <- read_dta(full_path)
 return(df)
}

lmb_data <- read_data("lmb-data.dta")

lmb_subset <- lmb_data %>%
 filter(lagdemvoteshare>.48 & lagdemvoteshare<.52)

lm_1 <- lm_robust(score ~ lagdemocrat, data = lmb_subset, clusters = id)
lm_2 <- lm_robust(score ~ democrat, data = lmb_subset, clusters = id)
lm_3 <- lm_robust(democrat ~ lagdemocrat, data = lmb_subset, clusters = id)

summary(lm_1)
summary(lm_2)
summary(lm_3)
```

Table 46. Replicated results based on ADA scores—close elections sample.

Dependent variable	ADA_{t+1}	ADA_t	DEM_{t+1}
Estimated gap	21.28*** (1.95)	47.71*** (1.36)	0.48*** (0.03)
N	915	915	915

Note: Cluster robust standard errors in parentheses. * p<0.10, ** p<0.05, *** p<0.01

We reproduce regression results from Lee, Moretti, and Butler in Table 46. While the results are close to Lee, Moretti, and Butler's original table, they are slightly different. But ignore that for now. The main thing to see is that we used regressions limited to the window right around the cutoff to estimate the effect. These are local regressions in the sense that they use data close to the cutoff. Notice the window we chose—we are only using observations between 0.48 and 0.52 vote share. So this regression is estimating the coefficient on D_t right around the cutoff. What happens if we use all the data?

STATA

lmb_2.do

```
1  * Use all the data
2  reg score lagdemocrat, cluster(id)
3  reg score democrat, cluster(id)
4  reg democrat lagdemocrat, cluster(id)
```

R

lmb_2.R

```
1  #using all data (note data used is lmb_data, not lmb_subset)
2
3  lm_1 <- lm_robust(score ~ lagdemocrat, data = lmb_data, clusters = id)
4  lm_2 <- lm_robust(score ~ democrat, data = lmb_data, clusters = id)
5  lm_3 <- lm_robust(democrat ~ lagdemocrat, data = lmb_data, clusters = id)
6
7  summary(lm_1)
8  summary(lm_2)
9  summary(lm_3)
```

Table 47. Results based on ADA scores—full sample.

Dependent variable	ADA_{t+1}	ADA_t	DEM_{t+1}
Estimated gap	31.50*** (0.48)	40.76*** (0.42)	0.82*** (0.01)
N	13,588	13,588	13,588

Note: Cluster robust standard errors in parentheses. * p<0.10, ** p<0.05, *** p<0.01

Notice that when we use all of the data, we get somewhat different effects (Table 47). The effect on future ADA scores gets larger by 10 points, but the contemporaneous effect gets smaller. The effect on incumbency, though, increases considerably. So here we see that

simply running the regression yields different estimates when we include data far from the cutoff itself.

Neither of these regressions included controls for the running variable though. It also doesn't use the recentering of the running variable. So let's do both. We will simply subtract 0.5 from the running variable so that values of 0 are where the vote share equals 0.5, negative values are Democratic vote shares less than 0.5, and positive values are Democratic vote shares above 0.5. To do this, type in the following lines:

STATA

lmb_3.do

```
1  * Re-center the running variable (voteshare)
2  gen demvoteshare_c = demvoteshare - 0.5
3  reg score lagdemocrat demvoteshare_c, cluster(id)
4  reg score democrat demvoteshare_c, cluster(id)
5  reg democrat lagdemocrat demvoteshare_c, cluster(id)
```

R

lmb_3.R

```
1  lmb_data <- lmb_data %>%
2    mutate(demvoteshare_c = demvoteshare - 0.5)
3
4  lm_1 <- lm_robust(score ~ lagdemocrat + demvoteshare_c, data = lmb_data,
   ↪    clusters = id)
5  lm_2 <- lm_robust(score ~ democrat + demvoteshare_c, data = lmb_data, clusters
   ↪    = id)
6  lm_3 <- lm_robust(democrat ~ lagdemocrat + demvoteshare_c, data = lmb_data,
   ↪    clusters = id)
7
8  summary(lm_1)
9  summary(lm_2)
10 summary(lm_3)
11
```

We report our analysis from the programming in Table 48. While the incumbency effect falls closer to what Lee et al. [2004] find, the

effects are still quite different.

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. To implement this in Stata, we can use the code shown in lmb_4.do.

Table 48. Results based on ADA scores—full sample.

Dependent variable	ADA_{t+1}	ADA_t	DEM_{t+1}
Estimated gap	33.45***	58.50***	0.55***
	(0.85)	(0.66)	(0.01)
N	13,577	13,577	13,577

Note: Cluster robust standard errors in parentheses. *$p < 0.10$. **$p < 0.05$. ***$p < 0.01$.

STATA

lmb_4.do

```
1  * Use all the data but interact the treatment variable with the running variable
2  xi: reg score i.lagdemocrat*demvoteshare_c, cluster(id)
3  xi: reg score i.democrat*demvoteshare_c, cluster(id)
4  xi: reg democrat i.lagdemocrat*demvoteshare_c, cluster(id)
```

R

lmb_4.R

```
1  lm_1 <- lm_robust(score ~ lagdemocrat*demvoteshare_c,
2                    data = lmb_data, clusters = id)
3  lm_2 <- lm_robust(score ~ democrat*demvoteshare_c,
4                    data = lmb_data, clusters = id)
5  lm_3 <- lm_robust(democrat ~ lagdemocrat*demvoteshare_c,
6                    data = lmb_data, clusters = id)
7
8  summary(lm_1)
9  summary(lm_2)
10 summary(lm_3)
11
```

In Table 49, we report the global regression analysis with the running variable interacted with the treatment variable. This pulled down the coefficients somewhat, but they remain larger than what was found when we used only those observations within 0.02 points of the 0.5. Finally, let's estimate the model with a quadratic.

STATA

lmb_5.do

```
1  * Use all the data but interact the treatment variable with the running variable and
   ↪  a quadratic
2  gen demvoteshare_sq = demvoteshare_c^2
3  xi: reg score lagdemocrat##c.(demvoteshare_c demvoteshare_sq), cluster(id)
4  xi: reg score democrat##c.(demvoteshare_c demvoteshare_sq), cluster(id)
5  xi: reg democrat lagdemocrat##c.(demvoteshare_c demvoteshare_sq),
   ↪  cluster(id)
```

R

lmb_5.R

```
lmb_data %>%
  mutate(demvoteshare_sq = demvoteshare_c^2)

lm_1 <- lm_robust(score ~ lagdemocrat*demvoteshare_c +
↪   lagdemocrat*demvoteshare_sq,
              data = lmb_data, clusters = id)
lm_2 <- lm_robust(score ~ democrat*demvoteshare_c +
↪   democrat*demvoteshare_sq,
              data = lmb_data, clusters = id)
lm_3 <- lm_robust(democrat ~ lagdemocrat*demvoteshare_c +
↪   lagdemocrat*demvoteshare_sq,
              data = lmb_data, clusters = id)

summary(lm_1)
summary(lm_2)
summary(lm_3)

```

Including the quadratic causes the estimated effect of a democratic victory on future voting to fall considerably (see Table 50). The effect on contemporaneous voting is smaller than what Lee et al. [2004] find, as is the incumbency effect. But the purpose here is simply to illustrate the standard steps using global regressions.

But notice, we are still estimating *global* regressions. And it is for that reason that the coefficient is larger. This suggests that there exist strong outliers in the data that are causing the distance at c_0 to spread more widely. So a natural solution is to again limit our analysis to a smaller window. What this does is drop the observations far away from c_0 and omit the influence of outliers from our estimation at the cutoff. Since we used +/−−0.02 last time, we'll use +/−−0.05 this time just to mix things up.

Table 49. Results based on ADA scores—full sample with linear interactions.

Dependent variable	ADA_{t+1}	ADA_t	DEM_{t+1}
Estimated gap	30.51***	55.43 ***	0.53***
	(0.82)	(0.64)	(0.01)
N	13,577	13,577	13,577

Note: Cluster robust standard errors in parentheses. *$p < 0.10$, **$p < 0.05$, ***$p < 0.01$

Table 50. Results based on ADA scores—full sample with linear and quadratic interactions.

Dependent variable	ADA_{t+1}	ADA_t	DEM_{t+1}
Estimated gap	13.03***	44.40 ***	0.32***
	(1.27)	(0.91)	(1.74)
N	13,577	13,577	13,577

Note: Cluster robust standard errors in parentheses. *$p < 0.10$, **$p < 0.05$, ***$p < 0.01$

STATA

lmb_6.do

```
1  * Use 5 points from the cutoff
2  xi: reg score lagdemocrat##c.(demvoteshare_c demvoteshare_sq) if
   ↪  lagdemvoteshare>.45 & lagdemvoteshare<.55, cluster(id)
3  xi: reg score democrat##c.(demvoteshare_c demvoteshare_sq) if
   ↪  lagdemvoteshare>.45 & lagdemvoteshare<.55, cluster(id)
4  xi: reg democrat lagdemocrat##c.(demvoteshare_c demvoteshare_sq) if
   ↪  lagdemvoteshare>.45 & lagdemvoteshare<.55, cluster(id)
```

R

lmb_6.R

```
1  lmb_data %>%
2    filter(demvoteshare > .45 & demvoteshare < .55) %>%
3    mutate(demvoteshare_sq = demvoteshare_c^2)
4
5  lm_1 <- lm_robust(score ~ lagdemocrat*demvoteshare_c +
   ↪  lagdemocrat*demvoteshare_sq,
6                    data = lmb_data, clusters = id)
7  lm_2 <- lm_robust(score ~ democrat*demvoteshare_c +
   ↪  democrat*demvoteshare_sq,
8                    data = lmb_data, clusters = id)
9  lm_3 <- lm_robust(democrat ~ lagdemocrat*demvoteshare_c +
   ↪  lagdemocrat*demvoteshare_sq,
10                   data = lmb_data, clusters = id)
11
12 summary(lm_1)
13 summary(lm_2)
14 summary(lm_3)
15
```

Table 51. Results based on ADA scores—close election sample with linear and quadratic interactions.

Dependent variable	ADA_{t+1}	ADA_t	DEM_{t+1}
Estimated gap	3.97*** (1.49)	46.88*** (1.54)	0.12*** (0.02)
N	2,441	2,441	2,441

Note: Cluster robust standard errors in parentheses. *$p < 0.10$, **$p < 0.05$, ***$p < 0.01$

As can be seen in Table 51, when we limit our analysis to +/− 0.05 around the cutoff, we are using more observations away from the cutoff than we used in our initial analysis. That's why we only have 2,441 observations for analysis as opposed to the 915 we had in our original analysis. But we also see that including the quadratic interaction pulled the estimated size on future voting down considerably, even when using the smaller sample.

But putting that aside, let's talk about all that we just did. First we fit a model without controlling for the running variable. But then we included the running variable, introduced in a variety of ways. For instance, we interacted the variable of Democratic vote share with the democratic dummy, as well as including a quadratic. In all this analysis, we extrapolated trends lines from the running variable beyond the support of the data to estimate local average treatment effects right at the cutoff.

But we also saw that the inclusion of the running variable in any form tended to reduce the effect of a victory for Democrats on future Democratic voting patterns, which was interesting. Lee et al. [2004] original estimate of around 21 is attenuated considerably when we include controls for the running variable, even when we go back to estimating very local flexible regressions. While the effect remains significant, it is considerably smaller, whereas the immediate effect remains quite large.

But there are still other ways to explore the impact of the treatment at the cutoff. For instance, while Hahn et al. [2001] clarified

assumptions about RDD—specifically, continuity of the conditional expected potential outcomes—they also framed estimation as a nonparametric problem and emphasized using local polynomial regressions. What exactly does this mean though in practice?

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 re-create important figures from Lee et al. [2004]. Figure 37 shows the relationship between the Democratic win (as a function of the running variable, Democratic vote share) and the candidates, second-period ADA score.

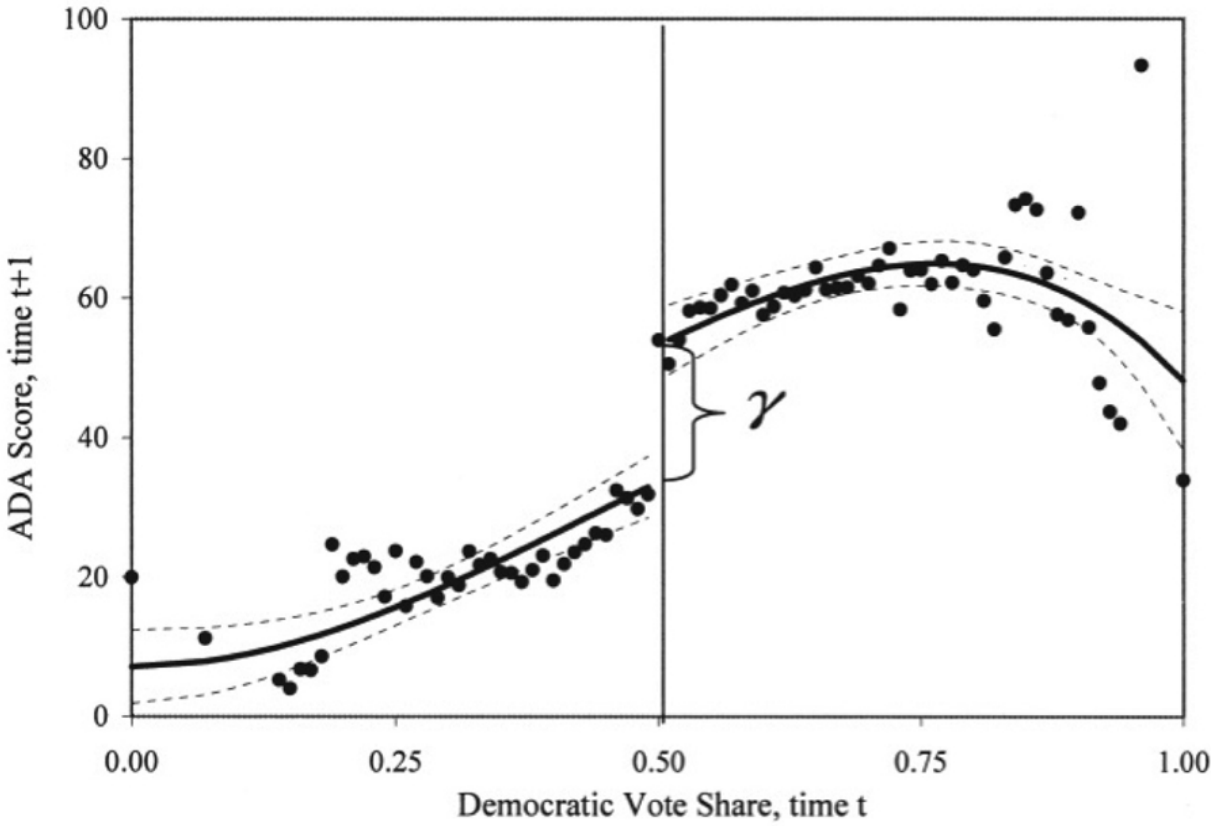


Figure 37. Showing total effect of initial win on future ADA scores. Reprinted from Lee, D. S., Moretti, E., and Butler, M. J. (2004). "Do Voters Affect or Elect Policies: Evidence from the U.S. House." *Quarterly Journal of Economics*, 119(3):807–859. Permission from Oxford University Press.

To reproduce this, there are a few options. You could manually create this figure yourself using either the "twoway" command in Stata or "ggplot" in R. But I'm going to show you using the canned cmogram routine that was created, as it's a quick-and-dirty way to get some information about the data.

STATA

lmb_7.do

```
1  * Nonparametric estimation graphic
2  ssc install cmogram
3  cmogram score lagdemvoteshare, cut(0.5) scatter line(0.5) qfitci
4  cmogram score lagdemvoteshare, cut(0.5) scatter line(0.5) lfit
5  cmogram score lagdemvoteshare, cut(0.5) scatter line(0.5) lowess
```

R

lmb_7.R

```
1  #aggregating the data
2  categories <- lmb_data$lagdemvoteshare
3
4  demmeans <- split(lmb_data$score, cut(lmb_data$lagdemvoteshare, 100)) %>%
5    lapply(mean) %>%
6    unlist()
7
8  agg_lmb_data <- data.frame(score = demmeans, lagdemvoteshare = seq(0.01,1,
   ↪    by = 0.01))
9
10 #plotting
11 lmb_data <- lmb_data %>%
12   mutate(gg_group = case_when(lagdemvoteshare > 0.5 ~ 1, TRUE ~ 0))
13
14 ggplot(lmb_data, aes(lagdemvoteshare, score)) +
15   geom_point(aes(x = lagdemvoteshare, y = score), data = agg_lmb_data) +
16   stat_smooth(aes(lagdemvoteshare, score, group = gg_group), method = "lm",
17               formula = y ~ x + I(x^2)) +
18   xlim(0,1) + ylim(0,100) +
19   geom_vline(xintercept = 0.5)
20
```

(continued)

R *(continued)*

```
ggplot(lmb_data, aes(lagdemvoteshare, score)) +
  geom_point(aes(x = lagdemvoteshare, y = score), data = agg_lmb_data) +
  stat_smooth(aes(lagdemvoteshare, score, group = gg_group), method = "loess")
  ↪  +
  xlim(0,1) + ylim(0,100) +
  geom_vline(xintercept = 0.5)

ggplot(lmb_data, aes(lagdemvoteshare, score)) +
  geom_point(aes(x = lagdemvoteshare, y = score), data = agg_lmb_data) +
  stat_smooth(aes(lagdemvoteshare, score, group = gg_group), method = "lm") +
  xlim(0,1) + ylim(0,100) +
  geom_vline(xintercept = 0.5)
```

Figure 38 shows the output from this program. Notice the similarities between what we produced here and what Lee et al. [2004] produced in their figure. The only differences are subtle changes in the binning used for the two figures.

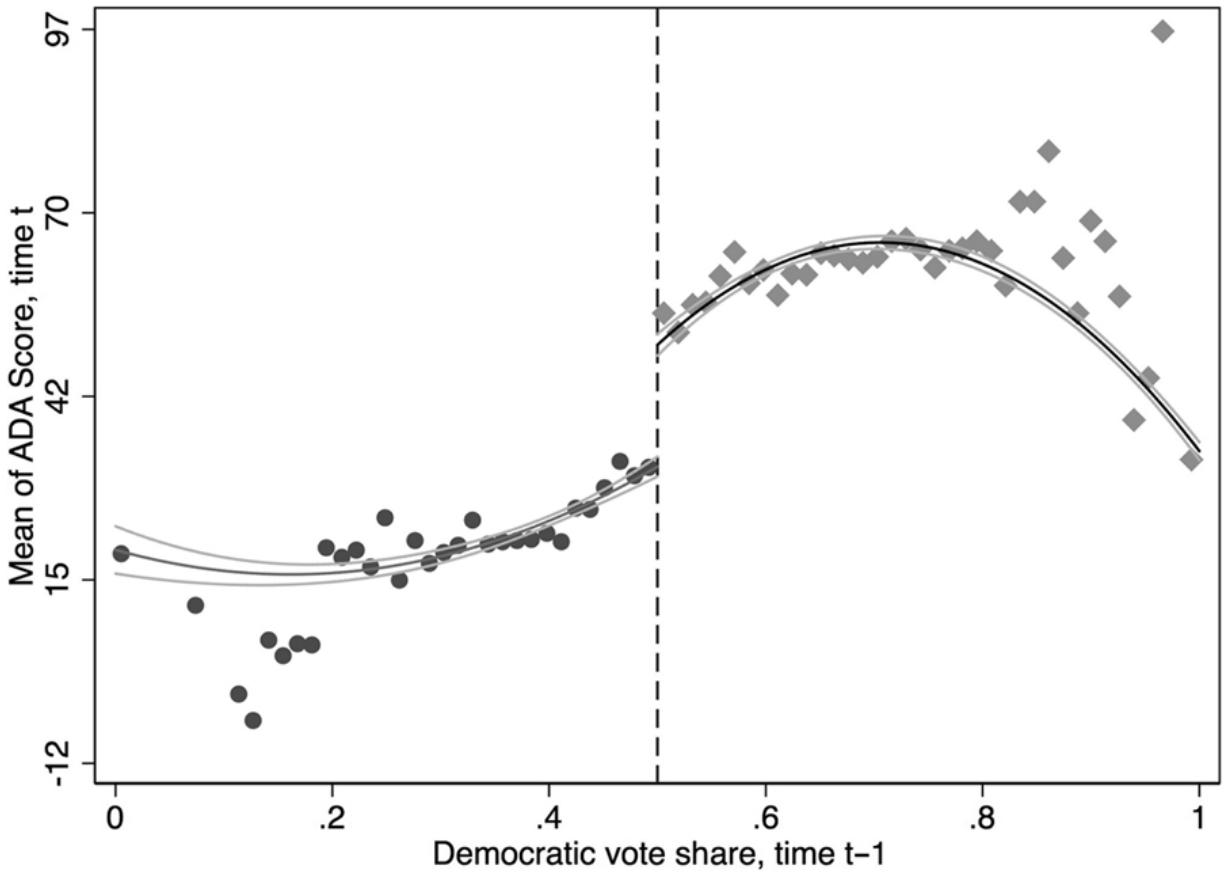


Figure 38. Using cmogram with quadratic fit and confidence intervals. Reprinted from Lee, D. S., Moretti, E., and Butler, M. J. (2004). “Do Voters Affect or Elect Policies: Evidence from the U.S. House.” *Quarterly Journal of Economics*, 119(3):807–859.

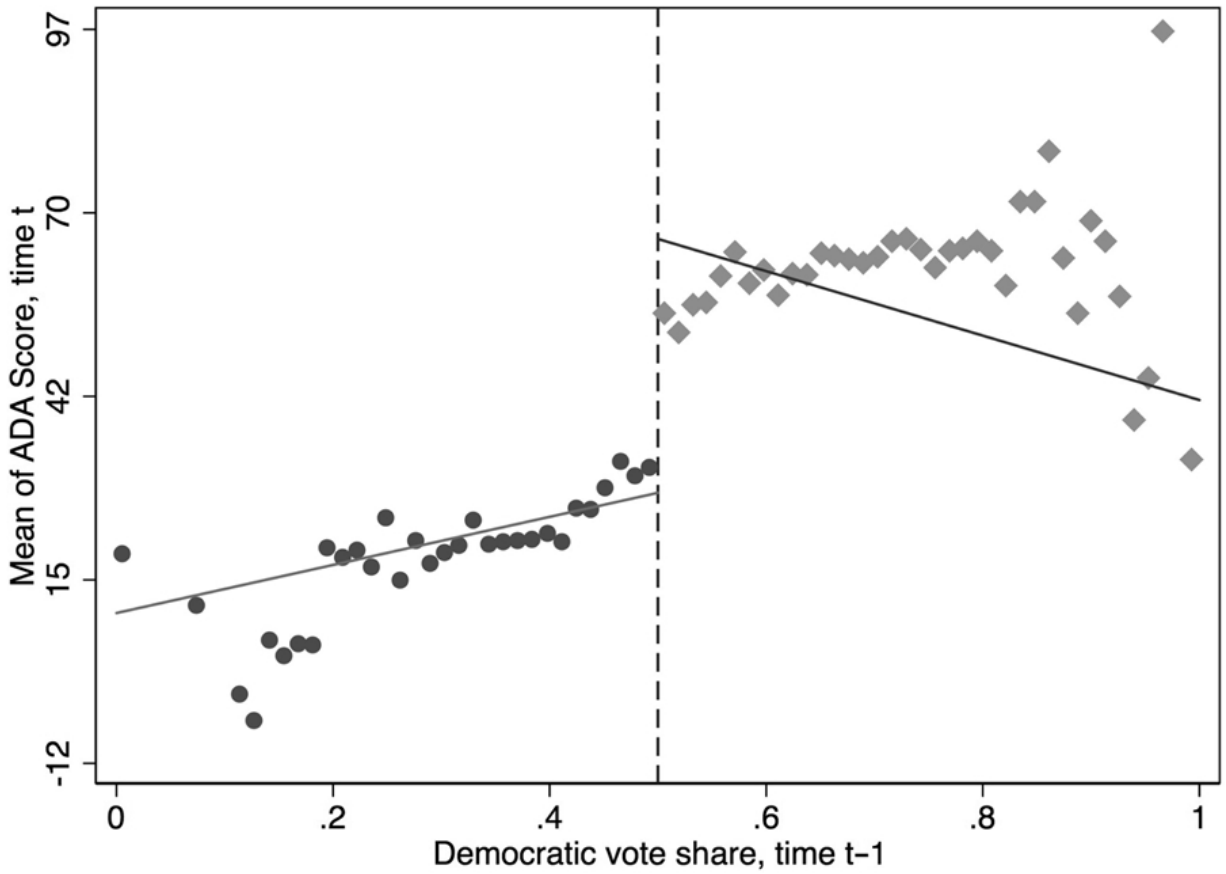


Figure 39. Using cmogram with linear fit. Reprinted from Lee, D. S., Moretti, E., and Butler, M. J. (2004). "Do Voters Affect or Elect Policies: Evidence from the U.S. House." *Quarterly Journal of Economics*, 119(3):807–859.

We have options other than a quadratic fit, though, and it's useful to compare this graph with one in which 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 simpler straight lines.

Figure 39 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. 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.

Finally, we can use a lowess fit. A lowess fit more or less crawls through the data and runs small regression on small cuts of data.

This can give the figure a zigzag appearance. We nonetheless show it in Figure 40.

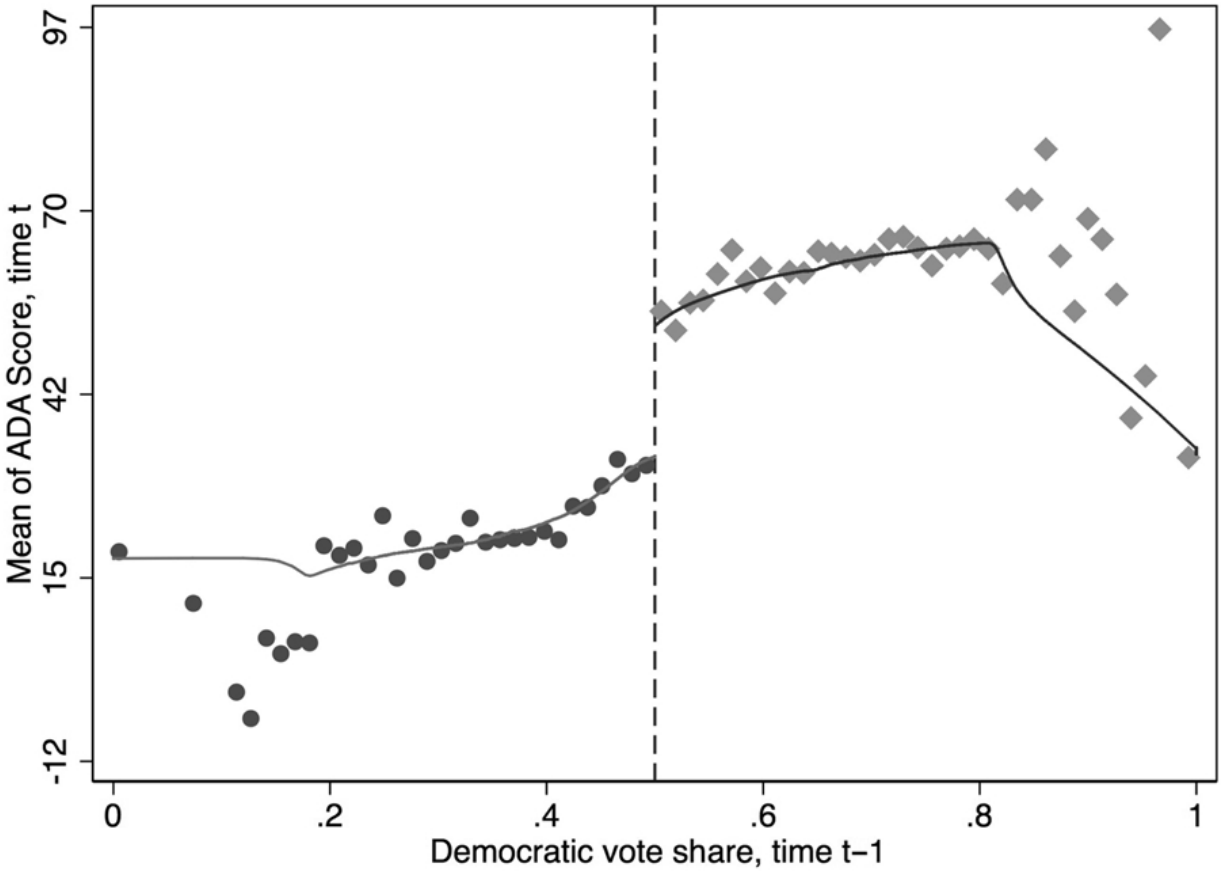


Figure 40. Using cmogram with lowess fit. Reprinted from Lee, D. S., Moretti, E., and Butler, M. J. (2004). "Do Voters Affect or Elect Policies: Evidence from the U.S. House." *Quarterly Journal of Economics*, 119(3):807–859.

If there don't appear to be any trends in the running variable, then the polynomials aren't going to buy you much. Some very good papers only report a linear fit because there weren't very strong trends to begin with. For instance, consider Carrell et al. [2011]. Those authors are interested in the causal effect of drinking on academic test outcomes for students at the Air Force Academy. 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 students' social life, there is a starker

increase in drinking at age 21 on its campus than might be the case for a more 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 presented a graph with only a linear fit. Your choice should be in large part based on what, to your eyeball, is the best fit of the data.

Hahn et al. [2001] have shown that one-sided kernel 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 using local linear nonparametric regressions instead. In these regressions, more weight is given to the observations at the center.

You can also estimate 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 the size of the bandwidth chosen. But in that sense, it's similar to what we've been doing. Figure 41 shows this visually.

STATA

lmb_8.do

```
1  * Note kernel-weighted local polynomial regression is a smoothing method.
2  capture drop sdem* x1 x0
3  lpoly score demvoteshare if democrat == 0, nograph kernel(triangle) gen(x0
   ↪  sdem0) bwidth(0.1)}
4  lpoly score demvoteshare if democrat == 1, nograph kernel(triangle) gen(x1
   ↪  sdem1)  bwidth(0.1)}
5  scatter sdem1 x1, color(red) msize(small) || scatter sdem0 x0, msize(small)
   ↪  color(red) xline(0.5,lstyle(dot)) legend(off) xtitle("Democratic vote share")
   ↪  ytitle("ADA score")
6
```

R

lmb_8.R

```
1  library(tidyverse)
2  library(stats)
3
4  smooth_dem0 <- lmb_data %>%
5    filter(democrat == 0) %>%
6    select(score, demvoteshare)
7  smooth_dem0 <- as_tibble(ksmooth(smooth_dem0$demvoteshare,
   ↪  smooth_dem0$score,
8                                   kernel = "box", bandwidth = 0.1))
```

(continued)

R *(continued)*

```
9
10
11  smooth_dem1 <- lmb_data %>%
12    filter(democrat == 1) %>%
13    select(score, demvoteshare) %>%
14    na.omit()
15  smooth_dem1 <- as_tibble(ksmooth(smooth_dem1$demvoteshare,
    ↪   smooth_dem1$score,
16                                   kernel = "box", bandwidth = 0.1))
17
18  ggplot() +
19    geom_smooth(aes(x, y), data = smooth_dem0) +
20    geom_smooth(aes(x, y), data = smooth_dem1) +
21    geom_vline(xintercept = 0.5)
22
23
24
25
```

A couple of final things. First, 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 always involve such placebos; even though they are not direct tests of the continuity assumption, they are indirect tests. Remember, when you are publishing, your readers aren't as familiar with this thing you're studying, so your task is explain to readers what you know. 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 the 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 that may differ left and right of the cutoff based on some bias-variance trade-off. Let's repeat our analysis using this nonparametric method. The coefficient is 46.48 with a standard error of 1.24.

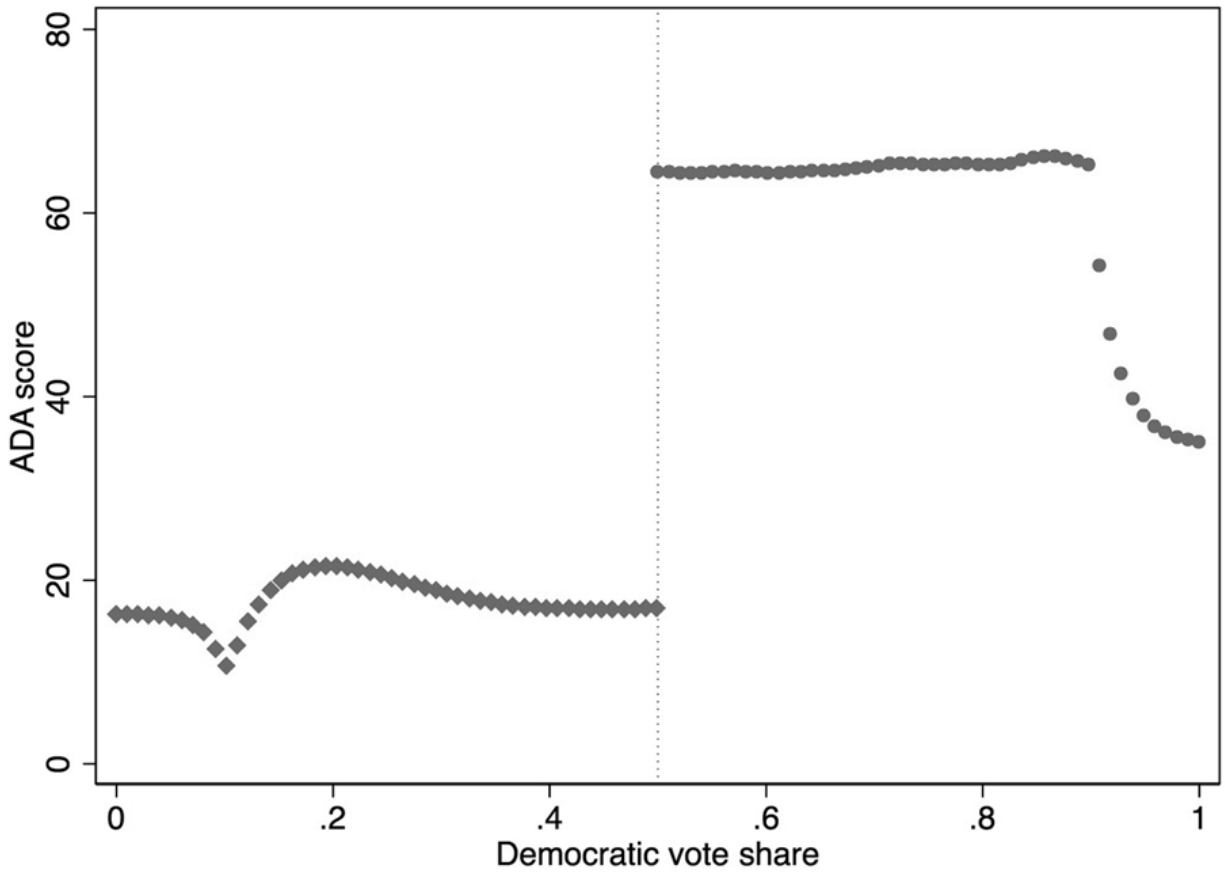


Figure 41. Local linear nonparametric regressions.

STATA

lmb_9.do

```
* Local polynomial point estimators with bias correction
ssc install rdrobust, replace
rdrobust score demvoteshare, c(0.5)
```

R

lmb_9.R

```
library(tidyverse)
library(rdrobust)

rdr <- rdrobust(y = lmb_data$score,
                x = lmb_data$demvoteshare, c = 0.5)
summary(rdr)
```

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 you may not have sufficient power to pick up this effect.

Finally, we look at the implementation of the McCrary density test. We will implement this test using local polynomial density estimation [Cattaneo et al., 2019]. This requires installing two files in Stata. Visually inspecting the graph in Figure 42, we see no signs that there was manipulation in the running variable at the cutoff.

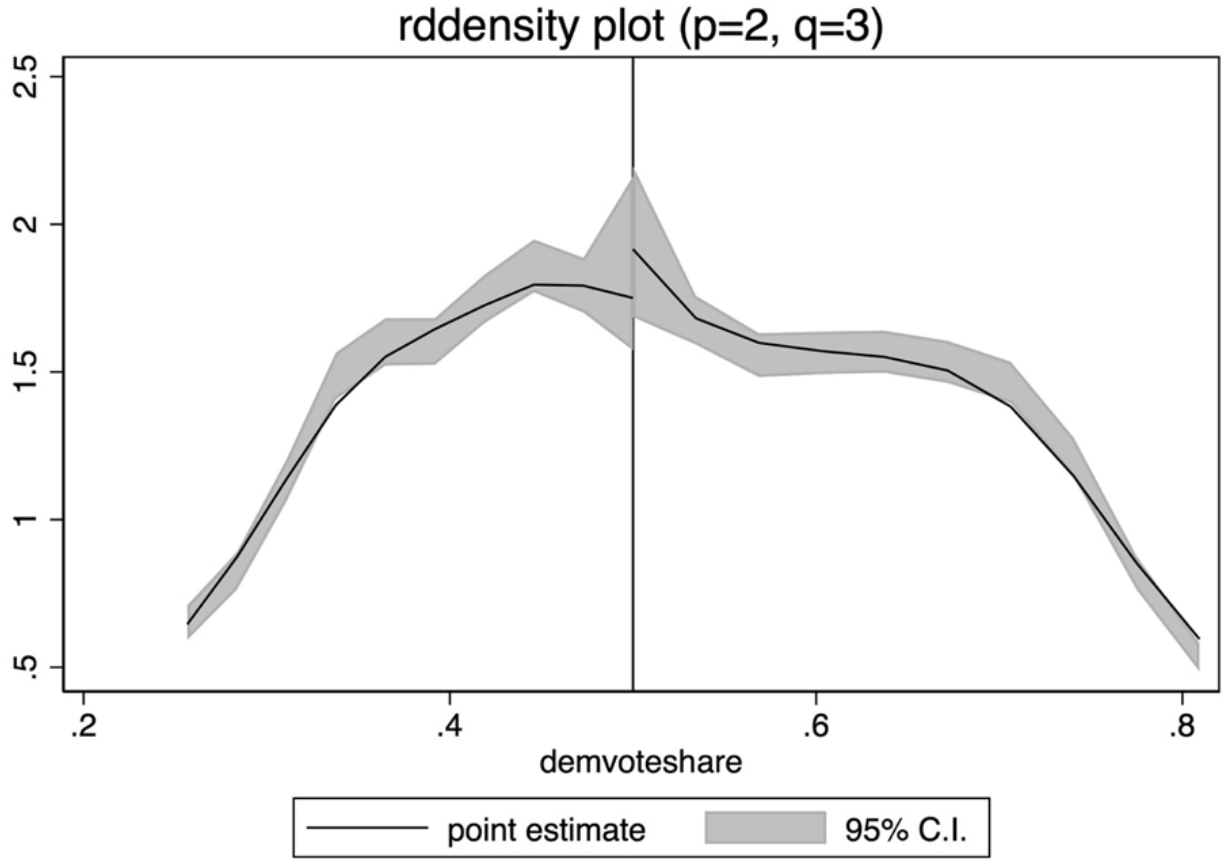


Figure 42. McCrary density test using local linear nonparametric regressions.

STATA

lmb_10.do

```
1  * McCrary density test
2  net install rddensity,
   ↪  from(https://sites.google.com/site/rdpackages/rddensity/stata) replace
3  net install lpdensity,
   ↪  from(https://sites.google.com/site/nppackages/lpdensity/stata) replace
4  rddensity demvoteshare, c(0.5) plot
```

R

lmb_10.R

```
1  library(tidyverse)
2  library(rddensity)
3  library(rdd)
4
5  DCdensity(lmb_data$demvoteshare, cutpoint = 0.5)
6
7  density <- rddensity(lmb_data$demvoteshare, c = 0.5)
8  rdplotdensity(density, lmb_data$demvoteshare)
```

Concluding remarks about close-election designs. Let's circle back to the close-election design. The design has since become practically a cottage industry within economics and political science. It has been extended to other types of elections and outcomes. One paper I like a lot used close gubernatorial elections to examine the effect of Democratic governors on the wage gap between workers of different races [Beland, 2015]. There are dozens more.

But a critique from Caughey and Sekhon [2011] called into question the validity of Lee's analysis on the House elections. They found that bare winners and bare losers in US House elections differed considerably on pretreatment covariates, which had not been formally evaluated by Lee et al. [2004]. And that covariate imbalance got even worse in the closest elections. Their conclusion is that the sorting problems got more severe, not less, in the closest of House races, suggesting that these races could not be used for an RDD.

At first glance, it appeared that this criticism by Caughey and Sekhon [2011] threw cold water on the entire close-election design, but we since know that is not the case. It appears that the Caughey and Sekhon [2011] criticism may have been only relevant for a subset of House races but did not characterize other time periods or other types of races. Eggers et al. [2014] evaluated 40,000 close elections, including the House in other time periods, mayoral races, and other types of races for political offices in the US and nine other countries. No other case that they encountered exhibited the type of pattern described by Caughey and Sekhon [2011]. Eggers et al. (2014) conclude that the assumptions behind RDD in the close-election design are likely to be met in a wide variety of electoral settings and is perhaps one of the best RD designs we have going forward.

Regression Kink Design

Many times, the concept of a running variable shifting a unit into treatment and in turn causing a jump in some outcome is sufficient. But there are some instances in which the idea of a "jump" doesn't describe what happens. A couple of papers by David Card and coauthors have extended the regression discontinuity design in order to handle these different types of situations. The most notable is Card et al. [2015], which introduced a new method called regression kink design, or RKD. The intuition is rather simple. Rather than the cutoff causing a discontinuous jump in the treatment variable at the cutoff, it changes the first derivative, which is known as a kink. Kinks are often embedded in policy rules, and thanks to Card et al. [2015], we can use kinks to identify the causal effect of a policy by exploiting the jump in the first derivative.

Card et al.'s [2015] paper applies the design to answer the question of whether the level of unemployment benefits affects the length of time spent unemployed in Austria. Unemployment benefits are based on income in a base period. There is then a minimum benefit level that isn't binding for people with low earnings. Then benefits are 55% of the earnings in the base period. There is a

maximum benefit level that is then adjusted every year, which creates a discontinuity in the schedule.

Figure 43 shows the relationship between base earnings and unemployment benefits around the discontinuity. There's a visible kink in the empirical relationship between average benefits and base earnings. You can see this in the sharp decline in the slope of the function as base-year earnings pass the threshold. Figure 44 presents a similar picture, but this time of unemployment duration. Again, there is a clear kink as base earnings pass the threshold. The authors conclude that increases in unemployment benefits in the Austrian context exert relatively large effects on unemployment duration.

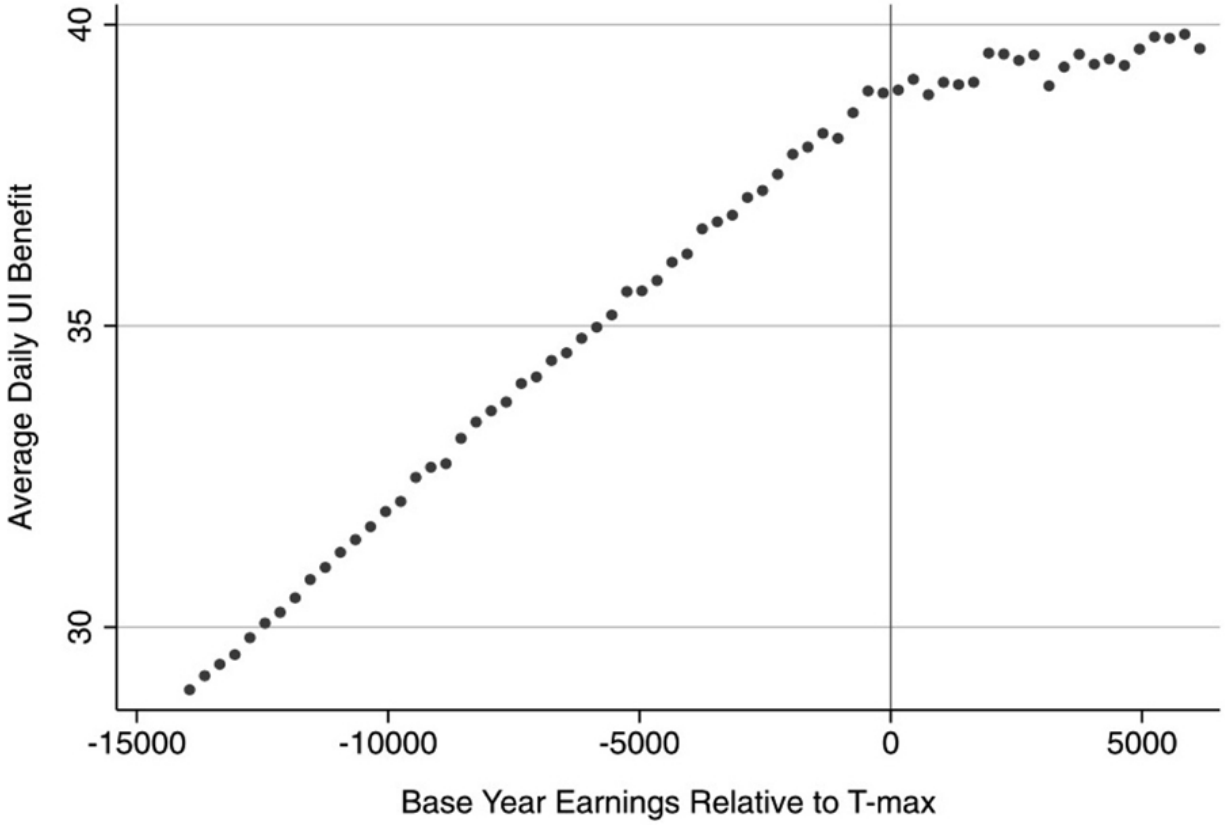


Figure 43. RKD kinks. Reprinted from Card, D., Lee, D. S., Pei, Z., and Weber, A. (2015). "Inference on Causal Effects in a Generalized Regression Kink Design." *Econometrica*, 84(6):2453–2483. Copyright ©2015 Wiley. Used with permission from John Wiley and Sons.

Conclusion

The regression discontinuity design is often considered a winning design because of its upside in credibly identifying causal effects. As with all designs, its credibility only comes from deep institutional knowledge, particularly surrounding the relationship between the running variable, the cutoff, treatment assignment, and the outcomes themselves. Insofar as one can easily find a situation in which a running variable passing some threshold leads to units being siphoned off into some treatment, then if continuity is believable, you're probably sitting on a great opportunity, assuming you can use it to do something theoretically interesting and policy relevant to others.

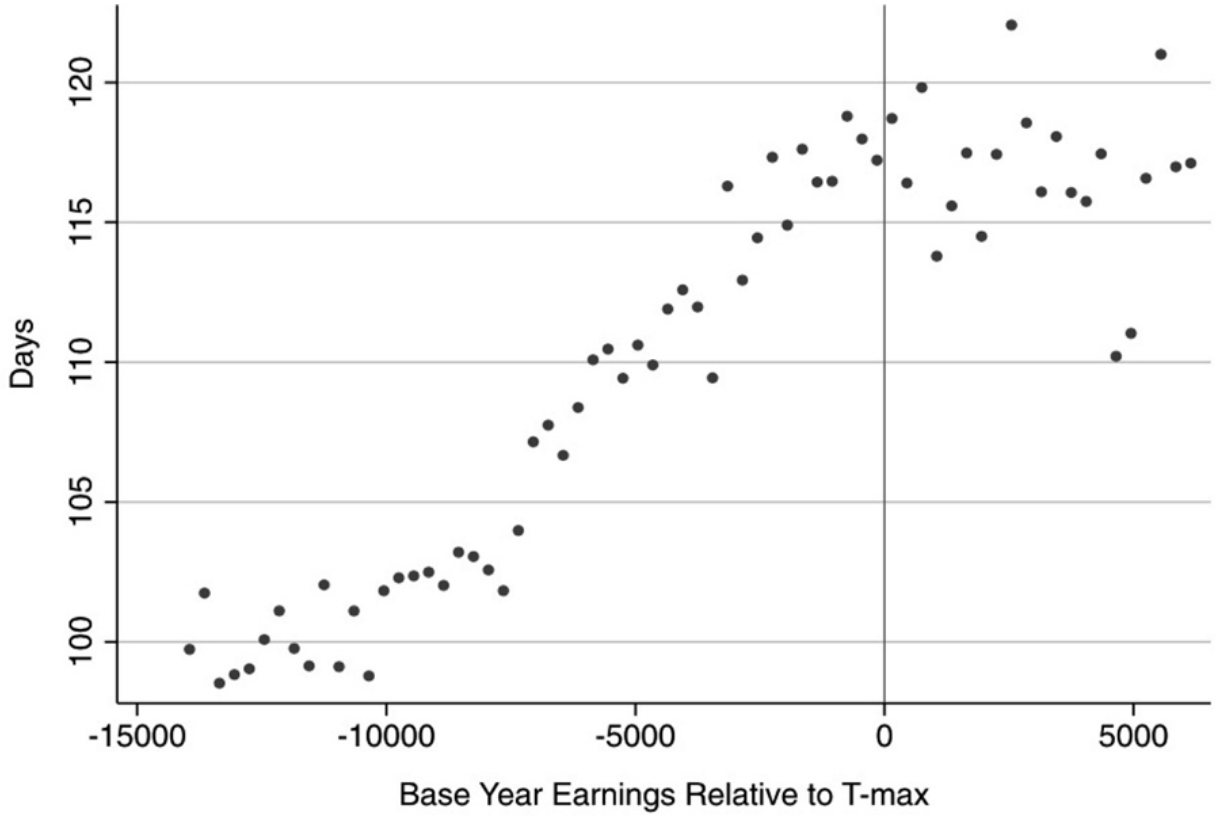


Figure 44. Unemployment duration. Reprinted from Card, D., Lee, D. S., Pei, Z., and Weber, A. (2015). "Inference on Causal Effects in a Generalized Regression Kink Design." *Econometrica*, 84(6):2453–2483. Copyright © 2015 Wiley. Used with permission from John Wiley and Sons.

Regression discontinuity design opportunities abound, particularly within firms and government agencies, for no other reason than that

these organizations face scarcity problems and must use some method to ration a treatment. Randomization is a fair way to do it, and that is often the method used. But a running variable is another method. Routinely, organizations will simply use a continuous score to assign treatments by arbitrarily picking a cutoff above which everyone receives the treatment. Finding these can yield a cheap yet powerfully informative natural experiment. This chapter attempted to lay out the basics of the design. But the area continues to grow at a lightning pace. So I encourage you to see this chapter as a starting point, not an ending point.

Notes

1 Thistlehwaite and Campbell [1960] 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. Knowing the treatment assignment allowed the authors to carefully estimate the causal effect of merit awards on future academic performance.

2 Hat tip to John Holbein for giving me these data.

3 Think about it for a moment. The backdoor criterion calculates differences in expected outcomes between treatment and control *for a given value of X*. But if the assignment variable only moves units into treatment when *X* passes some cutoff, then such calculations are impossible because there will not be units in treatment and control for any given value of *X*.

4 Mark Hoekstra is one of the more creative microeconomists I have met when it comes to devising compelling strategies for identifying causal effects in observational data, and this is one of my favorite papers by him.

5 "Don't be a jerk" applies even to situations when you aren't seeking proprietary data.

6 Van der Klaauw [2002] called the running variable the "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 here they mean the same thing.

7 Stata's poly command estimates kernel-weighted local polynomial regression.

8 RDHonest is available at https://github.com/kolesarm/RDHonest.

9 I discuss these assumptions and diagnostics in greater detail later in the chapter on instrument variables.

10 In those situations, anyway, where the treatment is desirable to the units.

11 https://sites.google.com/site/rdpackages/rddensity.

12 http://cran.r-project.org/web/packages/rdd/rdd.eps.

13 The honey badger doesn't care. It takes what it wants. See https://www.youtube.com/watch?v=4r7wHMg5Yjg.

Difference-in-Differences

What's the difference between me and you?
About five bank accounts, three ounces, and two vehicles.
Dr. Dre

The difference-in-differences design is an early quasi-experimental identification strategy for estimating causal effects that predates the randomized experiment by roughly eighty-five years. It has become the single most popular research design in the quantitative social sciences, and as such, it merits careful study by researchers everywhere.[1] In this chapter, I will explain this popular and important research design both in its simplest form, where a group of units is treated at the same time, and the more common form, where groups of units are treated at different points in time. My focus will be on the identifying assumptions needed for estimating treatment effects, including several practical tests and robustness exercises commonly performed, and I will point you to some of the work on difference-in-differences design (DD) being done at the frontier of research. I have included several replication exercises as well.

John Snow's Cholera Hypothesis

When thinking about situations in which a difference-in-differences design can be used, one usually tries to find an instance where a consequential treatment was given to some people or units but denied to others "haphazardly." This is sometimes called a "natural experiment" because it is based on naturally occurring variation in some treatment variable that affects only some units over time. All good difference-in-differences designs are based on some kind of natural experiment. And one of the most interesting natural experiments was also one of the first difference-in-differences designs. This is the story of how John Snow convinced the world that

cholera was transmitted by water, not air, using an ingenious natural experiment [Snow, 1855].

Cholera is a vicious disease that attacks victims suddenly, with acute symptoms such as vomiting and diarrhea. In the nineteenth century, it was usually fatal. There were three main epidemics that hit London, and like a tornado, they cut a path of devastation through the city. Snow, a physician, watched as tens of thousands suffered and died from a mysterious plague. Doctors could not help the victims because they were mistaken about the mechanism that caused cholera to spread between people.

The majority medical opinion about cholera transmission at that time was *miasma*, which said diseases were spread by microscopic poisonous particles that infected people by floating through the air. These particles were thought to be inanimate, and because microscopes at that time had incredibly poor resolution, it would be years before microorganisms would be seen. Treatments, therefore, tended to be designed to stop poisonous dirt from spreading through the air. But tried and true methods like quarantining the sick were strangely ineffective at slowing down this plague.

John Snow worked in London during these epidemics. Originally, Snow—like everyone—accepted the *miasma* theory and tried many ingenious approaches based on the theory to block these airborne poisons from reaching other people. He went so far as to cover the sick with burlap bags, for instance, but the disease still spread. People kept getting sick and dying. Faced with the theory's failure to explain cholera, he did what good scientists do—he changed his mind and began look for a new explanation.

Snow developed a novel theory about cholera in which the active agent was not an inanimate particle but was rather a living organism. This microorganism entered the body through food and drink, flowed through the alimentary canal where it multiplied and generated a poison that caused the body to expel water. With each evacuation, the organism passed out of the body and, importantly, flowed into England's water supply. People unknowingly drank contaminated water from the Thames River, which caused them to contract cholera. As they did, they would evacuate with vomit and diarrhea,

which would flow into the water supply again and again, leading to new infections across the city. This process repeated through a multiplier effect which was why cholera would hit the city in epidemic waves.

Snow's years of observing the clinical course of the disease led him to question the usefulness of *miasma* to explain cholera. While these were what we would call "anecdote," the numerous observations and imperfect studies nonetheless shaped his thinking. Here's just a few of the observations which puzzled him. He noticed that cholera transmission tended to follow human commerce. 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; he would not get sick if he remained docked. Cholera hit the poorest communities worst, and those people were the very same people who lived in the most crowded housing with the worst hygiene. He might find two apartment buildings next to one another, one would be heavily hit with cholera, but strangely the other one wouldn't. He then noticed that the first building would be contaminated by runoff from privies but the water supply in the second building was cleaner. While these observations weren't impossible to reconcile with *miasma*, they were definitely unusual and didn't seem obviously consistent with *miasmis*.

Snow tucked away more and more anecdotal evidence like these. But, while this evidence raised some doubts in his mind, he was not convinced. He needed a smoking gun if he were to eliminate all doubt that cholera was spread by water, not air. But where would he find that evidence? More importantly, what would evidence like that even look like?

Let's imagine the following thought experiment. If Snow was a dictator with unlimited wealth and power, how could he test his theory that cholera is waterborne? One thing he could do is flip a coin over each household member—heads you drink from the contaminated Thames, tails you drink from some uncontaminated source. Once the assignments had been made, Snow could simply compare cholera mortality between the two groups. If those who

drank the clean water were less likely to contract cholera, then this would suggest that cholera was waterborne.

Knowledge that physical randomization could be used to identify causal effects was still eighty-five years away. But there were other issues besides ignorance that kept Snow from physical randomization. Experiments like the one I just described are also impractical, infeasible, and maybe even unethical—which is why social scientists so often rely on natural experiments that mimic important elements of randomized experiments. But what natural experiment was there? Snow needed to find a situation where uncontaminated water had been distributed to a large number of people as if by random chance, and then calculate the difference between those those who did and did not drink contaminated water. Furthermore, the contaminated water would need to be allocated to people in ways that were unrelated to the ordinary determinants of cholera mortality, such as hygiene and poverty, implying a degree of balance on covariates between the groups. And then he remembered—a potential natural experiment in London a year earlier had reallocated clean water to citizens of London. Could this work?

In the 1800s, several water companies served different areas of the city. Some neighborhoods were even served by more than one company. They took their water from the Thames, which had been polluted by victims' evacuations via runoff. But in 1849, the Lambeth water company had moved its intake pipes upstream higher up the Thames, above the main sewage discharge point, thus giving its customers uncontaminated water. They did this to obtain cleaner water, but it had the added benefit of being too high up the Thames to be infected with cholera from the runoff. Snow seized on this opportunity. He realized that it had given him a natural experiment that would allow him to test his hypothesis that cholera was waterborne by comparing the households. If his theory was right, then the Lambeth houses should have lower cholera death rates than some other set of households whose water was infected with runoff—what we might call today the explicit counterfactual. He

found his explicit counterfactual in the Southwark and Vauxhall Waterworks Company.

Unlike Lambeth, the Southwark and Vauxhall Waterworks Company had *not* moved their intake point upstream, and Snow spent an entire book documenting similarities between the two companies' households. For instance, sometimes their service cut an irregular path through neighborhoods and houses such that the households on either side were very similar; the only difference being they drank different water with different levels of contamination from runoff. Insofar as the kinds of people that each company serviced were observationally equivalent, then perhaps they were similar on the relevant unobservables as well.

Snow meticulously collected data on household enrollment in water supply companies, going door to door asking household heads the name of their utility company. Sometimes these individuals didn't know, though, so he used a saline test to determine the source himself [Coleman, 2019]. He matched those data with the city's data on the cholera death rates at the household level. It was in many ways as advanced as any study we might see today for how he carefully collected, prepared, and linked a variety of data sources to show the relationship between water purity and mortality. But he also displayed scientific ingenuity for how he carefully framed the research question and how long he remained skeptical until the research design's results convinced him otherwise. After combining everthing, he was able to generate extremely persuasive evidence that influenced policymakers in the city.[2]

Snow wrote up all of his analysis in a manuscript entitled *On the Mode of Communication of Cholera* [Snow, 1855]. Snow's main evidence was striking, and I will discuss results based on Table XII and Table IX (not shown) in Table 69. The main difference between my version and his version of Table XII is that I will use his data to estimate a treatment effect using difference-in-differences.

Table XII. In 1849, there were 135 cases of cholera per 10,000 households at Southwark and Vauxhall and 85 for Lambeth. But in

1854, there were 147 per 100,000 in Southwark and Vauxhall, whereas Lambeth's cholera cases per 10,000 households fell to 19.

Table 69. Modified Table XII (Snow 1854).

Company name	1849	1854
Southwark and Vauxhall	135	147
Lambeth	85	19

While Snow did not explicitly calculate the difference-in-differences, the ability to do so was there [Coleman, 2019]. If we difference Lambeth's 1854 value from its 1849 value, followed by the same after and before differencing for Southwark and Vauxhall, we can calculate an estimate of the ATT equaling 78 fewer deaths per 10,000. While Snow would go on to produce evidence showing cholera deaths were concentrated around a pump on Broad Street contaminated with cholera, he allegedly considered the simple difference-in-differences the more convincing test of his hypothesis.

The importance of the work Snow undertook to understand the causes of cholera in London cannot be overstated. It not only lifted our ability to estimate causal effects with observational data, it advanced science and ultimately saved lives. Of Snow's work on the cause of cholera transmission, Freedman [1991] states:

> The force of [Snow's] 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. [298]

Estimation

A simple table. 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.[3] Assume that the intervention is clean water, which I'll write as D, and our objective is to estimate D's causal effect on cholera deaths. Let cholera deaths be represented by the variable Y. Can we identify the causal effect of D if we just compare the post-treatment 1854 Lambeth cholera death values to that of the 1854 Southwark and Vauxhall values? This is in many ways an obvious choice, and in fact, it is one of the more common naive approaches to causal inference. After all, we have a control group, don't we? Why can't we just compare a treatment group to a control group? Let's look and see.

Table 70. Compared to what? Different companies.

Company	Outcome
Lambeth	$Y = L + D$
Southwark and Vauxhall	$Y = SV$

Table 71. Compared to what? Before and after.

Company	Time	Outcome
Lambeth	Before	$Y = L$
	After	$Y = L + (T + D)$

One of the things we immediately must remember is that the simple difference in outcomes, which is all we are doing here, only collapsed to the ATE if the treatment had been randomized. But it is never randomized in the real world where most choices if not all choices made by real people is endogenous to potential outcomes. Let's represent now the differences between Lambeth and Southwark and Vauxhall with fixed level differences, or fixed effects, represented by L and SV. Both are unobserved, unique to each company, and fixed over time. What these fixed effects mean is that even if Lambeth hadn't changed its water source there, would still be

something determining cholera deaths, which is just the time-invariant unique differences between the two companies as it relates to cholera deaths in 1854.

Table 72. Compared to what? Difference in each company's differences.

Companies	Time	Outcome	D_1	D_2
Lambeth	Before	$Y = L$		
	After	$Y = L + T + D$	$T + D$	
				D
Southwark and Vauxhall	Before	$Y = SV$		
	After	$Y = SV + T$	T	

When we make a simple comparison between Lambeth and Southwark and Vauxhall, we get an estimated causal effect equalling $D + (L - SV)$. Notice the second term, $L - SV$. We've seen this before. It's the selection bias we found from the decomposition of the simple difference in outcomes from earlier in the book.

Okay, so say we realize that we cannot simply make cross-sectional comparisons between two units because of selection bias. Surely, though, we can compare a unit to itself? This is sometimes called an interrupted time series. Let's consider that simple before-and-after difference for Lambeth now.

While this procedure successfully eliminates the Lambeth fixed effect (unlike the cross-sectional difference), it doesn't give me an unbiased estimate of D because differences can't eliminate the natural changes in the cholera deaths over time. Recall, these events were oscillating in waves. I can't compare Lambeth before and after (T+D) because of T, which is an omitted variable.

The intuition of the DD strategy is remarkably simple: combine these two simpler approaches so the selection bias and the effect of time are, in turns, eliminated. Let's look at it in the following table.

The first difference, D_1, does the simple before-and-after 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 *D*.

But there's a a key assumption with a DD design, and that assumption is discernible even in this table. We are assuming that there is no time-variant company specific unobservables. Nothing unobserved in Lambeth households that is changing between these two periods that *also* determines cholera deaths. This is equivalent to assuming that *T* is the same for all units. And we call this the *parallel trends* assumption. We will discuss this assumption repeatedly as the chapter proceeds, as it is the most important assumption in the design's engine. If you can buy off on the parallel trends assumption, then DD will identify the causal effect.

DD is a powerful, yet amazingly simple design. Using repeated observations on a treatment and control unit (usually several units), we can eliminate the unobserved heterogeneity to provide a credible estimate of the average treatment effect on the treated (ATT) by transforming the data in very specific ways. But when and why does this process yield the correct answer? Turns out, there is more to it than meets the eye. And it is imperative on the front end that you understand what's under the hood so that you can avoid conceptual errors about this design.

The simple 2×2 *DD*. The cholera case is a particular kind of DD design that Goodman-Bacon [2019] calls the 2 ×2 DD design. The 2×2 DD design has a treatment group *k* and untreated group *U*. There is a pre-period for the treatment group, pre*(k)*; a post-period for the treatment group, post*(k)*; a pre-treatment period for the untreated group, pre*(U)*; and a post-period for the untreated group, post*(U)* So:

$$\widehat{\delta}_{kU}^{2\times2} = \left(\overline{y}_k^{\text{post}(k)} - \overline{y}_k^{\text{pre}(k)}\right) - \left(\overline{y}_U^{\text{post}(k)} - \overline{y}_U^{\text{pre}(k)}\right)$$

where $\widehat{\delta}_{kU}$ is the estimated ATT for group *k*, and $\bar{y}$ is the sample mean for that particular group in a particular time period. The first paragraph differences the treatment group, *k*, after minus before, the

second paragraph differences the untreated group, U, after minus before. And once those quantities are obtained, we difference the second term from the first.

But this is simply the mechanics of calculations. What exactly is this estimated parameter mapping onto? To understand that, we must convert these sample averages into conditional expectations of potential outcomes. But that is easy to do when working with sample averages, as we will see here. First let's rewrite this as a conditional expectation.

$$\widehat{\delta}^{2\times2}_{kU} = \Big(E[Y_k \mid \text{Post}] - E[Y_k \mid \text{Pre}]\Big) - \Big(E[Y_U \mid \text{Post}] - E[Y_U \mid \text{Pre}]\Big)$$

Now let's use the switching equation, which transforms historical quantities of Y into potential outcomes. As we've done before, we'll do a little trick where we add zero to the right-hand side so that we can use those terms to help illustrate something important.

$$\widehat{\delta}^{2\times2}_{kU} = \underbrace{\Big(E[Y^1_k \mid \text{Post}] - E[Y^0_k \mid \text{Pre}]\Big) - \Big(E[Y^0_U \mid \text{Post}] - E[Y^0_U \mid \text{Pre}]\Big)}_{\text{Switching equation}}$$
$$+ \underbrace{E[Y^0_k \mid \text{Post}] - E[Y^0_k \mid \text{Post}]}_{\text{Adding zero}}$$

Now we simply rearrange these terms to get the decomposition of the 2×2 DD in terms of conditional expected potential outcomes.

$$\widehat{\delta}^{2\times2}_{kU} = \underbrace{E[Y^1_k \mid \text{Post}] - E[Y^0_k \mid \text{Post}]}_{\text{ATT}}$$
$$+ \underbrace{\Big[E[Y^0_k \mid \text{Post}] - E[Y^0_k \mid \text{Pre}]\Big] - \Big[E[Y^0_U \mid \text{Post}] - E[Y^0_U \mid \text{Pre}]\Big]}_{\text{Non-parallel trends bias in } 2\times2 \text{ case}}$$

Now, let's study this last term closely. This simple 2×2 difference-in-differences will isolate the ATT (the first term) if and only if the

second term zeroes out. But why would this second term be zero? It would equal zero if the first difference involving the treatment group, k, equaled the second difference involving the untreated group, U.

But notice the term in the second line. Notice anything strange about it? The object of interest is Y^0, which is some outcome in a world without the treatment. But it's the *post* period, and in the post period, $Y = Y^1$ not Y^0 by the switching equation. Thus, the first term is *counterfactual*. And as we've said over and over, counterfactuals are not observable. This bottom line is often called the parallel trends assumption and it is by definition untestable since we cannot observe this counterfactual conditional expectation. We will return to this again, but for now I simply present it for your consideration. *DD and the Minimum Wage*. Now I'd like to talk about more explicit economic content, and the minimum wage is as good a topic as any. The modern use of DD was brought into the social sciences through esteemed labor economist Orley Ashenfelter [1978]. His study was no doubt influential to his advisee, David Card, arguably the greatest labor economist of his generation. Card would go on to use the method in several pioneering studies, such as Card [1990]. But I will focus on one in particular—his now-classicminimum wage study [Card and Krueger, 1994].

Card and Krueger [1994] is an infamous study both because of its use of an explicit counterfactual for estimation, and because the study challenges many people's common beliefs about the negative effects of the minimum wage. It lionized a massive back-and-forth minimum-wage literature that continues to this day.[4] So controversial was this study that James Buchanan, the Nobel Prize winner, called those influenced by Card and Krueger [1994] "camp following whores" in a letter to the editor of the *Wall Street Journal* [Buchanan, 1996].[5]

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 in labor markets characterized by monopsony, minimum wages can

increase employment. Therefore, there are strong theoretical reasons to believe that the effect of the minimum wage on employment is ultimately an empirical question depending on many local contextual factors. This is where Card and Krueger [1994] entered. Could they uncover whether minimum wages were ultimately harmful or helpful in some local economy?

It's always useful to start these questions with a simple thought experiment: if you had a billion dollars, complete discretion and could run a randomized experiment, how would you test whether minimum wages increased or decreased employment? You might go across the hundreds of local labor markets in the United States and flip a coin—heads, you raise the minimum wage; tails, you keep it at the status quo. As we've done before, these kinds of thought experiments are useful for clarifying both the research design and the causal question.

Lacking a randomized experiment, Card and Krueger [1994] decided on a next-best solution by comparing two neighboring states before and after a minimum-wage increase. It was essentially the same strategy that Snow used in his cholera study and a strategy that economists continue to use, in one form or another, to this day [Dube et al., 2010].

New Jersey was set to experience an increase in the state minimum wage from $4.25 to $5.05 in November 1992, but neighboring Pennsylvania's minimum wage was staying at $4.25. Realizing they had an opportunity to evaluate the effect of the minimum-wage increase by comparing the two states before and after, they fielded a survey of about four hundred fast-food restaurants in both states—once in February 1992 (before) and again in November (after). The responses from this survey were then used to measure the outcomes they cared about (i.e., employment). As we saw with Snow, we see again here that shoe leather is as important as any statistical technique in causal inference.

Let's look at whether the minimum-wage hike in New Jersey in fact raised the minimum wage by examining the distribution of wages in the fast food stores they surveyed. Figure 54 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.

As a caveat, notice how effective this is at convincing the reader that the minimum wage in New Jersey was binding. This piece of data visualization is not a trivial, or even optional, strategy to be taken in studies such as this. Even John Snow presented carefully designed maps of the distribution of cholera deaths throughout London. Beautiful pictures displaying the “first stage” effect of the intervention on the treatment are crucial in the rhetoric of causal inference, and few have done it as well as Card and Krueger.

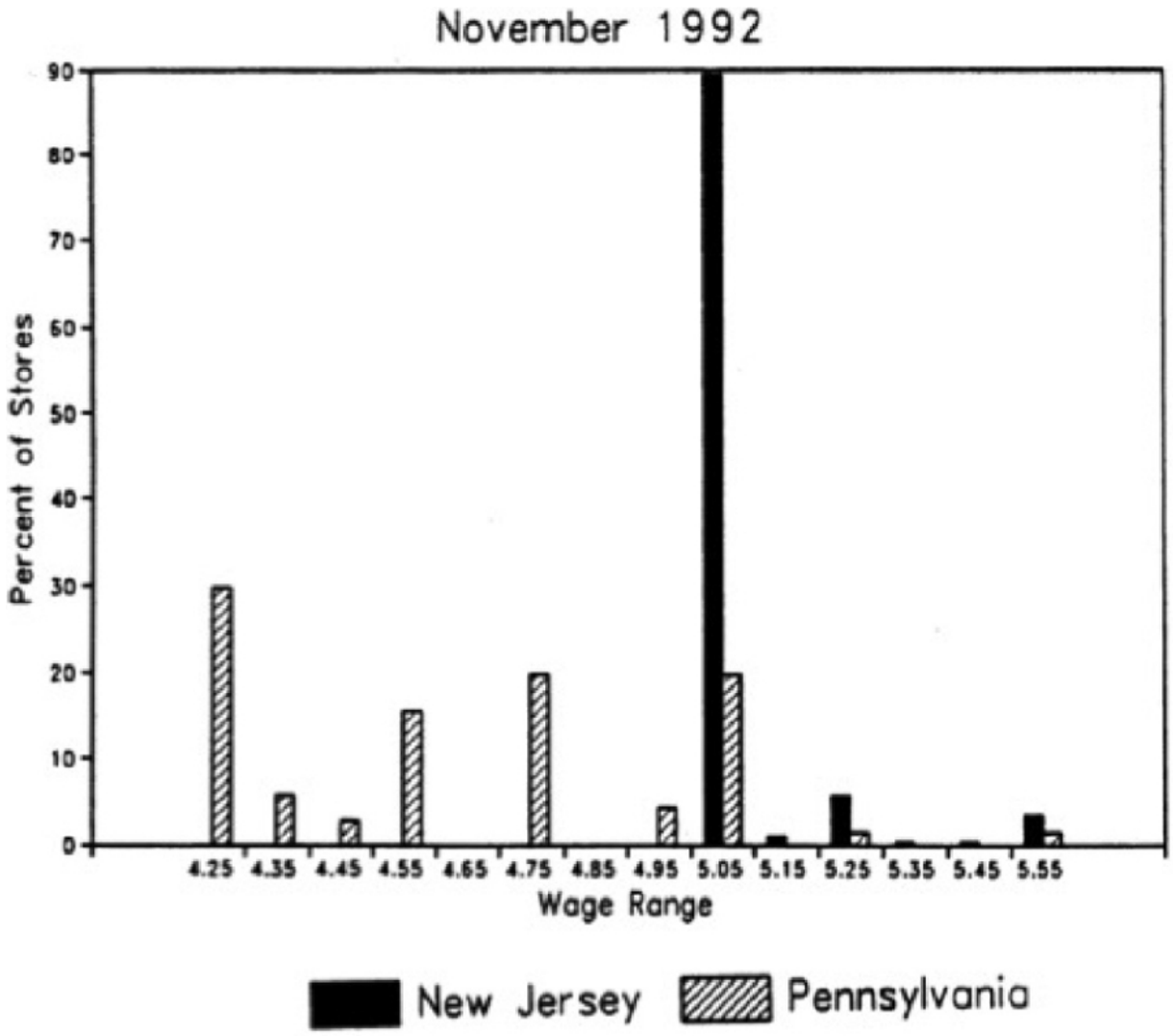


Figure 54. Distribution of wages for NJ and PA in November 1992. Reprinted from Card, D. and Krueger, A. (1994). "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania." *American Economic Review*, 84:772–793. Reprinted with permission from authors.

Let's remind ourselves what we're after—the average causal effect of the minimum-wage hike on employment, or the ATT. Using our decomposition of the 2×2 DD from earlier, we can write it out as:

$$\widehat{\delta}^{2\times2}_{NJ,PA} = \underbrace{E[Y^1_{NJ} \mid \text{Post}] - E[Y^0_{NJ} \mid \text{Post}]}_{\text{ATT}}$$
$$+ \underbrace{\Big[E[Y^0_{NJ} \mid \text{Post}] - E[Y^0_{NJ} \mid \text{Pre}]\Big] - \Big[E[Y^0_{PA} \mid \text{Post}] - E[Y^0_{PA} \mid \text{Pre}]\Big]}_{\text{Non-parallel trends bias}}$$

Again, we see the key assumption: the parallel-trends assumption, which is represented by the first difference in the second line. Insofar as parallel trends holds in this situation, then the second term goes to zero, and the 2×2 DD collapses to the ATT.

The 2×2 DD requires differencing employment in NJ and PA, then differencing those first differences. This set of steps estimates the true ATT so long as the parallel-trends bias is zero. When that is true, $\widehat{\delta}^{2\times2}$ is equal to δ_{ATT}. If this bottom line is not zero, though, then simple 2×2 suffers from unknown bias—could bias it upwards, could bias it downwards, could flip the sign entirely. Table 73 shows the results of this exercise from Card and Krueger [1994].

Table 73. Simple DD using sample averages on full-time employment.

	Stores by state		
Dependent variable	**PA**	**NJ**	**NJ − PA**
FTW before	23.3	20.44	−2.89
	(1.35)	(0.51)	(1.44)
FTE after	21.147	21.03	−0.14
	(0.94)	(0.52)	(1.07)
Change in mean FTE	−2.16	0.59	2.76
	(1.25)	(0.54)	(1.36)

Note: Standard errors in parentheses.

Here you see the result that surprised many people. Card and Krueger [1994] estimate an ATT of +2.76 additional mean full-time-equivalent employment, as opposed to some negative value which would be consistent with competitive input markets. Herein we get

Buchanan's frustration with the paper, which is based mainly on a particular model he had in mind, rather than a criticism of the research design the authors used.

While differences in sample averages will identify the ATT under the parallel assumption, we may want to use multivariate regression instead. For instance, if you need to avoid omitted variable bias through controlling for endogenous covariates that vary over time, then you may want to use regression. Such strategies are another way of saying that you will need to close some known critical backdoor. Another reason for the equation is that by controlling for more appropriate covariates, you can reduce residual variance and improve the precision of your DD estimate.

Using the switching equation, and assuming a constant state fixed effect and time fixed effect, we can write out a simple regression model estimating the causal effect of the minimum wage on employment, *Y*.

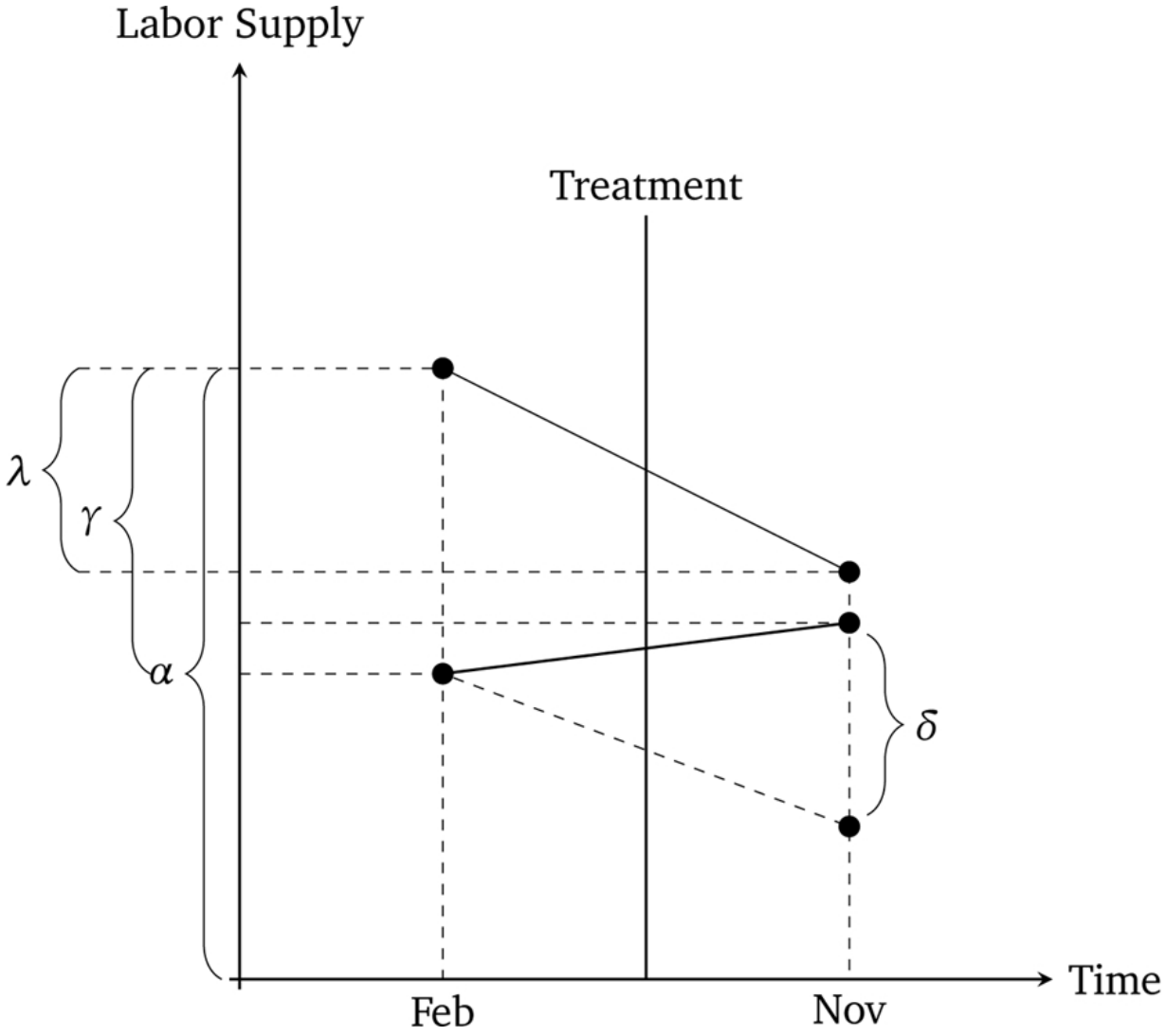


Figure 55. DD regression diagram.

This simple 2×2 is estimated with the following equation:

$$Y_{its} = \alpha + \gamma\, NJ_s + \lambda D_t + \delta(NJ \times D)_{st} + \varepsilon_{its}$$

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, which I will list in order according to setting the dummies equal to one and/or zero:

1. PA Pre: α
2. PA Post: $\alpha + \lambda$
3. NJ Pre: $\alpha + \gamma$

4. NJ Post: $\alpha + \gamma + \lambda + \delta$

We can visualize the 2×2 DD parameter in Figure 55.

Now before we hammer the parallel trends assumption for the billionth time, I wanted to point something out here which is a bit subtle. But do you see the δ parameter floating in the air above the November line in the Figure 55? This is the difference between a counterfactual level of employment (the bottom black circle in November on the negatively sloped dashed line) and the actual level of employment (the above black circle in November on the positively sloped solid line) for New Jersey. It is therefore the ATT, because the ATT is equal to

$$\delta = E[Y^1_{NJ,\text{Post}}] - E[Y^0_{NJ,\text{Post}}]$$

wherein the first is observed (because $Y = Y^1$ in the post period) and the latter is unobserved for the same reason.

Now here's the kicker: OLS will always estimate that δ line *even if the counterfactual slope had been something else*. That's because OLS uses Pennsylvania's change over time to project a point starting at New Jersey's pre-treatment value. When OLS has filled in that missing amount, the parameter estimate is equal to the difference between the observed post-treatment value and that projected value based on the slope of Pennsylvania *regardless of whether that Pennsylvania slope was the correct benchmark for measuring New Jersey's counterfactual slope*. OLS always estimates an effect size using the slope of the untreated group as the counterfactual, regardless of whether that slope is in fact the correct one.

But, see what happens when Pennsylvania's slope is equal to New Jersey's counterfactual slope? Then that Pennsylvania slope used in regression will mechanically estimate the ATT. In other words, only when the Pennsylvania slope is the counterfactual slope for New Jersey will OLS coincidentally identify that true effect. Let's see that here in Figure 56.

Notice the two δ listed: on the left is the true parameter δ_{ATT}. On the right is the one estimated by OLS, $\widehat{\delta}_{OLS}$. The falling solid line is

the observed Pennsylvania change, whereas the falling solid line labeled “observed NJ” is the change in observed employment for New Jersey between the two periods.

The true causal effect, δ_{ATT}, is the line from the “observed NJ” point and the “counterfactual NJ” point. But OLS does not estimate this line. Instead, OLS uses the falling Pennsylvania line to draw a parallel line from the February NJ point, which is shown in thin gray. And OLS simply estimates the vertical line from the observed NJ point to the postNJ point, which as can be seen underestimates the true causal effect.

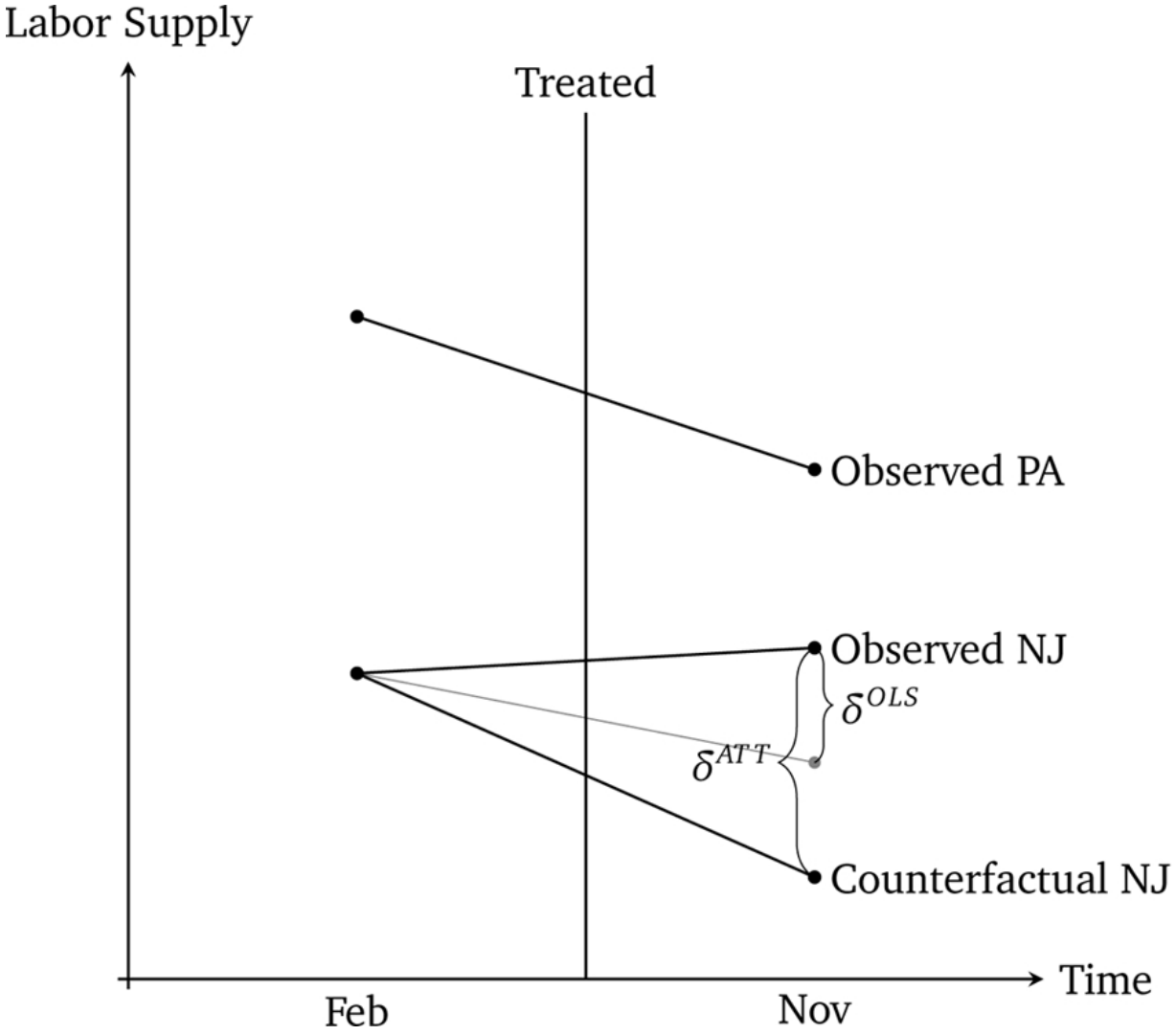


Figure 56. DD regression diagram without parallel trends.

Here we see the importance of the parallel trends assumption. The only situation under which the OLS estimate equals the ATT is when the counterfactual NJ just coincidentally lined up with the gray OLS line, which is a line parallel to the slope of the Pennsylvania line. Herein lies the source of understandable skepticism of many who have been paying attention: why should we base estimation on this belief in a coincidence? After all, this is a counterfactual trend, and therefore it is unobserved, given it never occurred. Maybe the counterfactual would've been the gray line, but maybe it would've been some other unknown line. It could've been anything—we just don't know.

This is why I like to tell people that the parallel trends assumption is actually just a restatement of the strict exogeneity assumption we discussed in the panel chapter. What we are saying when we appeal to parallel trends is that we have found a control group who approximates the traveling path of the treatment group *and* that the treatment is not endogenous. If it is endogenous, then parallel trends is always violated because in counterfactual the treatment group would've diverged anyway, regardless of the treatment.

Before we see the number of tests that economists have devised to provide some reasonable confidence in the belief of the parallel trends, I'd like to quickly talk about standard errors in a DD design.

Inference

Many studies employing DD strategies use data from many years—not just one pre-treatment and one 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 understate the standard deviation of the estimators, and so standard errors are biased downward, "too small," and therefore overreject the

null hypothesis. Bertrand et al. [2004] propose the following solutions:

1. Block bootstrapping standard errors.
2. Aggregating the data into one pre and one post period.
3. Clustering standard errors at the group level.

Block bootstrapping. If the block is a state, then you simply sample states with replacement for bootstrapping. Block bootstrap is straightforward and only requires a little programming involving loops and storing the estimates. As the mechanics are similar to that of randomization inference, I leave it to the reader to think about how they might tackle this.

Aggregation. This approach ignores the time-series dimensions altogether, and if there is only one pre and post period and one untreated group, it's as simple as it sounds. You simply average the groups into one pre and post period, and conduct difference-in-differences on those aggregated. But if you have differential timing, it's a bit unusual because you will need to partial out state and year fixed effects before turning the analysis into an analysis involving residualization. Essentially, for those common situations where you have multiple treatment time periods (which we discuss later in greater detail), you would regress the outcome onto panel unit and time fixed effects and any covariates. You'd then obtain the residuals for only the treatment group. You then divide the residuals only into a pre and post period; you are essentially at this point ignoring the never-treated groups. And then you regress the residuals on the after dummy. It's a strange procedure, and does not recover the original point estimate, so I focus instead on the third.

Clustering. Correct treatment of standard errors sometimes makes the number of groups very small: in Card and Krueger [1994], the number of groups is only two. More common than not, researchers will use the third option (clustering the standard errors by group). I have only one time seen someone do all three of these; it's rare

though. Most people will present just the clustering solution—most likely because it requires minimal programming.

For clustering, there is no programming required, as most software packages allow for it already. You simply adjust standard errors by clustering at the group level, as we discussed in the earlier chapter, or the level of treatment. For state-level panels, that would mean clustering at the state level, which allows for arbitrary serial correlation in errors within a state over time. This is the most common solution employed.

Inference in a panel setting is independently an interesting area. When the number of clusters is small, then simple solutions like clustering the standard errors no longer suffice because of a growing false positive problem. In the extreme case with only one treatment unit, the over-rejection rate at a significance of 5% can be as high as 80% in simulations even using the wild bootstrap technique which has been suggested for smaller numbers of clusters [Cameron et al., 2008; MacKinnon andWebb, 2017]. In such extreme cases where there is only one treatment group, I have preferred to use randomization inference following Buchmueller et al. [2011].

Providing Evidence for Parallel Trends Through Event Studies and Parallel Leads

A redundant rant about parallel pre-treatment DD coefficients (because I'm worried one was not enough). Given the critical importance of the parallel trends assumption in identifying causal effects with the DD design, and given that one of the observations needed to evaluate the parallel-trends assumption is not available to the researcher, one might throw up their hands in despair. But economists are stubborn, and they have spent decades devising ways to test whether it's reasonable to believe in parallel trends. We now discuss the obligatory test for any DD design—the event study. Let's rewrite the decomposition of the 2×2 DD again.

$$\widehat{\delta}_{kU}^{2\times 2} = \underbrace{E[Y_k^1 \mid \text{Post}] - E[Y_k^0 \mid \text{Post}]}_{\text{ATT}}$$
$$+ \underbrace{\Big[E[Y_k^0 \mid \text{Post}] - E[Y_k^0 \mid \text{Pre}]\Big] - \Big[E[Y_U^0 \mid \text{Post}] - E[Y_U^0 \mid \text{Pre}]\Big]}_{\text{Non-parallel trends bias}}$$

We are interested in the first term, ATT, but it is contaminated by selection bias when the second term does not equal zero. Since evaluating the second term requires the counterfactual, $E[Y_k^0 \mid Post]$, we are unable to do so directly. What economists typically do, instead, is compare placebo pre-treatment leads of the DD coefficient. If DD coefficients in the pre-treatment periods are statistically zero, then the difference-in-differences between treatment and control groups followed a similar trend prior to treatment. And here's the rhetorical art of the design: *if* they had been similar before, *then* why wouldn't they continue to be post-treatment?

But notice that this rhetoric is a kind of proof by assertion. Just because they were similar before does not logically require they be the same after. Assuming that the future is like the past is a form of the gambler's fallacy called the "reverse position." Just because a coin came up heads three times in a row does not mean it will come up heads the fourth time—not without further assumptions. Likewise, we are not obligated to believe that that counterfactual trends would be the same post-treatment because they had been similar pretreatment without further assumptions about the predictive power of pre-treatment trends. But to make such assumptions is again to make untestable assumptions, and so we are back where we started.

One situation where parallel trends would be obviously violated is if the treatment itself was endogenous. In such a scenario, the assignment of the treatment status would be directly dependent on potential outcomes, and absent the treatment, potential outcomes would've changed regardless. Such traditional endogeneity requires more than merely lazy visualizations of parallel leads. While the test

is important, technically pre-treatment similarities are neither necessary nor sufficient to guarantee parallel counterfactual trends [Kahn-Lang and Lang, 2019]. The assumption is not so easily proven. You can never stop being diligent in attempting to determine whether groups of units endogenously selected into treatment, the presence of omitted variable biases, various sources of selection bias, and open backdoor paths. When the structural error term in a dynamic regression model is uncorrelated with the treatment variable, you have strict exogeneity, and that is what gives you parallel trends, and that is what makes you able to make meaningful statements about your estimates.

Checking the pre-treatment balance between treatment and control groups. Now with that pessimism out of the way, let's discuss event study plots because though they are not direct tests of the parallel trends assumption, they have their place because they show that the two groups of units were comparable on dynamics in the pre-treatment period.[6] Such conditional independence concepts have been used profitably throughout this book, and we do so again now.

Authors have tried showing the differences between treatment and control groups a few different ways. One way is to simply show the raw data, which you can do if you have a set of groups who received the treatment at the same point in time. Then you would just visually inspect whether the pre-treatment dynamics of the treatment group differed from that of the control group units.

But what if you do not have a single treatment date? What if instead you have differential timing wherein groups of units adopt the treatment at different points? Then the concept of pre-treatment becomes complex. If New Jersey raised its minimum wage in 1992 and New York raised its minimum wage in 1994, but Pennsylvania never raised its minimum wage, the pre-treatment period is defined for New Jersey (1991) and New York (1993), but not Pennsylvania. Thus, how do we go about testing for pre-treatment differences in that case? People have done it in a variety of ways.

One possibility is to plot the raw data, year by year, and simply eyeball. You would compare the treatment group with the never-

treated, for instance, which might require a lot of graphs and may also be awkward looking. Cheng and Hoekstra [2013] took this route, and created a separate graph comparing treatment groups with an untreated group for each different year of treatment. The advantage is its transparent display of the raw unadjusted data. No funny business. The disadvantage of this several-fold. First, it may be cumbersome when the number of treatment groups is large, making it practically impossible. Second, it may not be beautiful. But third, this necessarily assumes that the only control group is the never-treated group, which in fact is not true given what Goodman-Bacon [2019] has shown. Any DD is a combination of a comparison between the treatment and the never treated, an early treated compared to a late treated, and a late treated compared to an early treated. Thus only showing the comparison with the never treated is actually a misleading presentation of the underlying mechanization of identification using an twoway fixed-effects model with differential timing.

Anderson et al. [2013] took an alternative, creative approach to show the comparability of states with legalized medical marijuana and states without. As I said, the concept of a pre-treatment period for a control state is undefined when pre-treatment is always in reference to a specific treatment date which varies across groups. So, the authors construct a recentered time path of traffic fatality rates for the control states by assigning random treatment dates to all control counties and then plotting the average traffic fatality rates for each group in years leading up to treatment and beyond. This approach has a few advantages. First, it plots the raw data, rather than coefficients from a regression (as we will see next). Second, it plots that data against controls. But its weakness is that technically, the control series is not in fact *true*. It is chosen so as to give a comparison, but when regressions are eventually run, it will not be based on this series. But the main main shortcoming is that technically it is not displaying any of the control groups that will be used for estimation Goodman-Bacon [2019]. It is not displaying a comparison between the treated and the never treated; it is not a comparison between the early and late treated; it is not a

comparison between the late and early treated. While a creative attempt to evaluate the pre-treatment differences in leads, it does not in fact technically show that.

The current way in which authors evaluate the pre-treatment dynamics between a treatment and control group with differential timing is to estimate a regression model that includes treatment leads and lags. I find that it is always useful to teach these concepts in the context of an actual paper, so let's review an interesting working paper by Miller et al. [2019].

Affordable Care Act, expanding Medicaid and population mortality. A provocative new study by Miller et al. [2019] examined the expansion of Medicaid under the Affordable Care Act. They were primarily interested in the effect that this expansion had on population mortality. Earlier work had cast doubt on Medicaid's effect on mortality [Baicker et al., 2013; Finkelstein et al., 2012], so revisiting the question with a larger sample size had value.

Like Snow before them, the authors link data sets on deaths with a large-scale federal survey data, thus showing that shoe leather often goes hand in hand with good design. They use these data to evaluate the causal impact of Medicaid enrollment on mortality using a DD design. Their focus is on the near-elderly adults in states with and without the Affordable Care Act Medicaid expansions and they find a 0.13-percentage-point decline in annual mortality, which is a 9.3% reduction over the sample mean, as a result of the ACA expansion. This effect is a result of a reduction in disease-related deaths and gets larger over time. Medicaid, in this estimation, saved a non-trivial number of lives.

As with many contemporary DD designs, Miller et al. [2019] evaluate the pre-treatment leads instead of plotting the raw data by treatment and control. Post-estimation, they plotted regression coefficients with 95% confidence intervals on their treatment leads and lags. Including leads and lags into the DD model allowed the reader to check both the degree to which the post-treatment treatment effects were dynamic, and whether the two groups were

comparable on outcome dynamics pre-treatment. Models like this one usually follow a form like:

$$Y_{its} = \gamma_s + \lambda_t + \sum_{\tau=-q}^{-1} \gamma_\tau D_{s\tau} + \sum_{\tau=0}^{m} \delta_\tau D_{s\tau} + x_{ist} + \varepsilon_{ist}$$

Treatment occurs in year 0. You include q leads or anticipatory effects and m lags or post-treatment effects.

Miller et al. [2019] produce four event studies that when taken together tell the main parts of the story of their paper. This is, quite frankly, the art of the rhetoric of causal inference—visualization of key estimates, such as "first stages" as well as outcomes and placebos. The event study plots are so powerfully persuasive, they will make you a bit jealous, since oftentimes yours won't be nearly so nice. Let's look at the first three. State expansion of Medicaid under the Affordable Care Act increased Medicaid eligibility (Figure 57), which is not altogether surprising. It also caused an increase in Medicaid coverage (Figure 58), and as a consequence reduced the percentage of the uninsured population (Figure 59). All three of these are simply showing that the ACA Medicaid expansion had "bite"—people enrolled and became insured.

There are several features of these event studies that should catch your eye. First, look at Figure 57. The pre-treatment coefficients are nearly on the zero line itself. Not only are they nearly zero in their point estimate, but their standard errors are very small. This means these are very precisely estimated zero differences between individuals in the two groups of states prior to the expansion.

The second thing you see, though, is the elephant in the room. Post-treatment, the probability that someone becomes eligible for Medicaid immediately shoots up to 0.4 and while not as precise as the pre-treatment coefficients, the authors can rule out effects as low as 0.3 to 0.35. These are large increases in eligibility, and the fact that the coefficients prior to the treatment are basically zero, we find it easy to believe that the risen coefficients post-treatment were caused by the ACA's expansion of Medicaid in states.

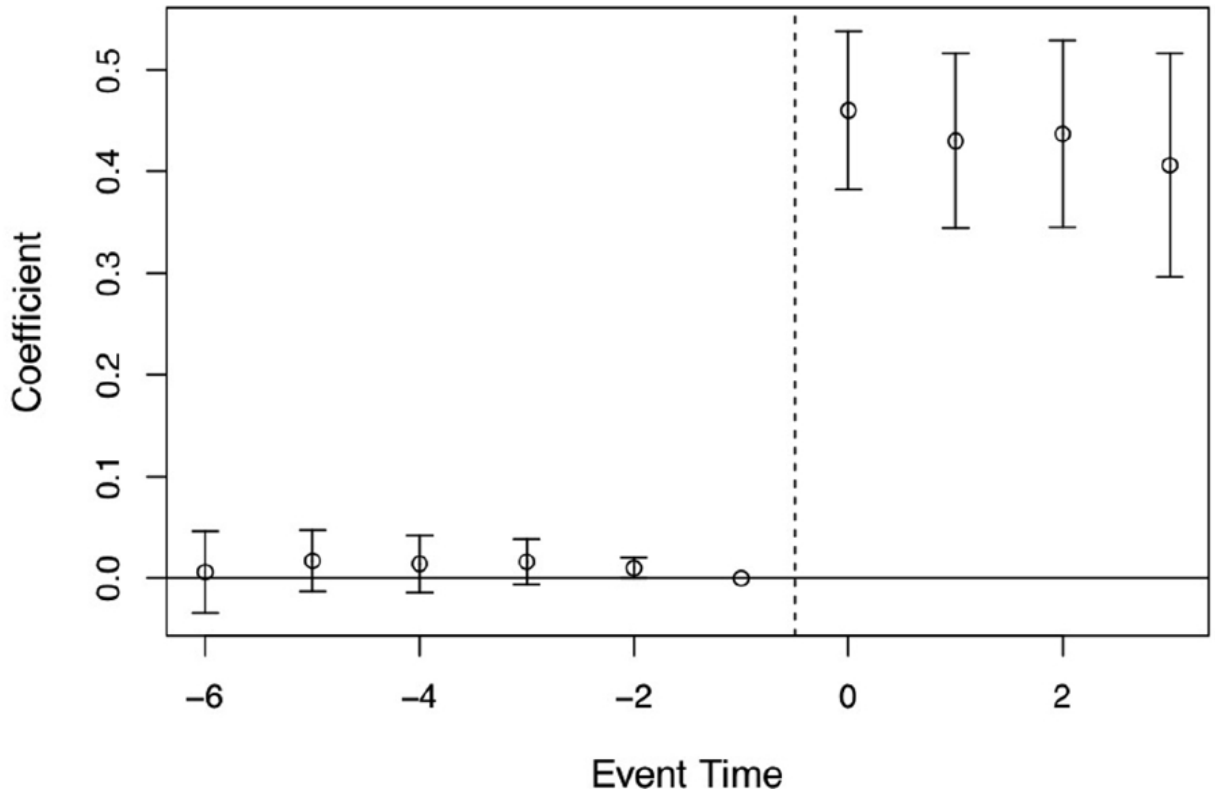


Figure 57. Estimates of Medicaid expansion's effects on eligibility using leads and lags in an event-study model. Miller, S., Altekruse, S., Johnson, N., and Wherry, L. R. (2019). Medicaid and mortality: New evidence from linked survey and administrative data. Working Paper No. 6081, National Bureau of Economic Research, Cambridge, MA. Reprinted with permission from authors.

Of course, I would not be me if I did not say that *technically* the zeroes pre-treatment do not therefore mean that the post-treatment difference between counterfactual trends and observed trends are zero, but doesn't it seem compelling when you see it? Doesn't it compel you, just a little bit, that the changes in enrollment and insurance status were probably caused by the Medicaid expansion? I daresay a table of coefficients with leads, lags, and standard errors would probably not be as compelling even though it is the identical information. Also, it is only fair that the skeptic refuse these patterns with new evidence of what it is other than the Medicaid expansion. It is not enough to merely hand wave a criticism of omitted variable bias; the critic must be as engaged in this phenomenon as the authors themselves, which is how empiricists earn the right to critique someone else's work.

Similar graphs are shown for coverage—prior to treatment, the two groups of individuals in treatment and control were similar with regards to their coverage and uninsured rate. But post-treatment, they diverge dramatically. Taken together, we have the "first stage," which means we can see that the Medicaid expansion under the ACA had "bite." Had the authors failed to find changes in eligibility, coverage, or uninsured rates, then any evidence from the secondary outcomes would have doubt built in. This is the reason it is so important that you examine the first stage (treatment's effect on usage), as well as the second stage (treatment's effect on the outcomes of interest).

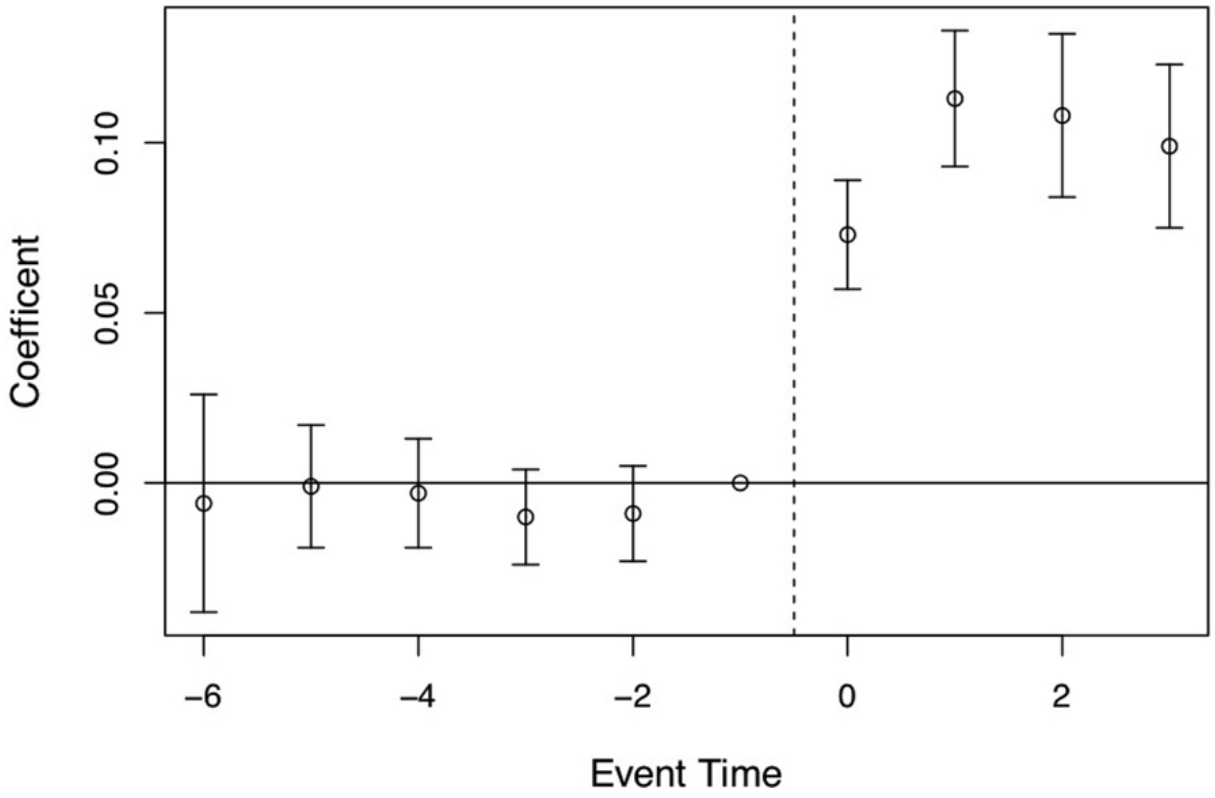


Figure 58. Estimates of Medicaid expansion's effects on coverage using leads and lags in an event-study model. Miller, S., Altekruse, S., Johnson, N., and Wherry, L. R. (2019). Medicaid and mortality: New evidence from linked survey and administrative data. Working Paper No. 6081, National Bureau of Economic Research, Cambridge, MA. Reprinted with permission from authors.

But now let's look at the main result—what effect did this have on population mortality itself? Recall, Miller et al. [2019] linked

administrative death records with a large-scale federal survey. So they actually know who is on Medicaid and who is not. John Snow would be proud of this design, the meticulous collection of high-quality data, and all the shoeleather the authors showed.

This event study is presented in Figure 60. A graph like this is the contemporary heart and soul of a DD design, both because it conveys key information regarding the comparability of the treatment and control groups in their dynamics just prior to treatment, and because such strong data visualization of main effects are powerfully persuasive. It's quite clear looking at it that there was no difference between the trending tendencies of the two sets of state prior to treatment, making the subsequent divergence all the more striking.

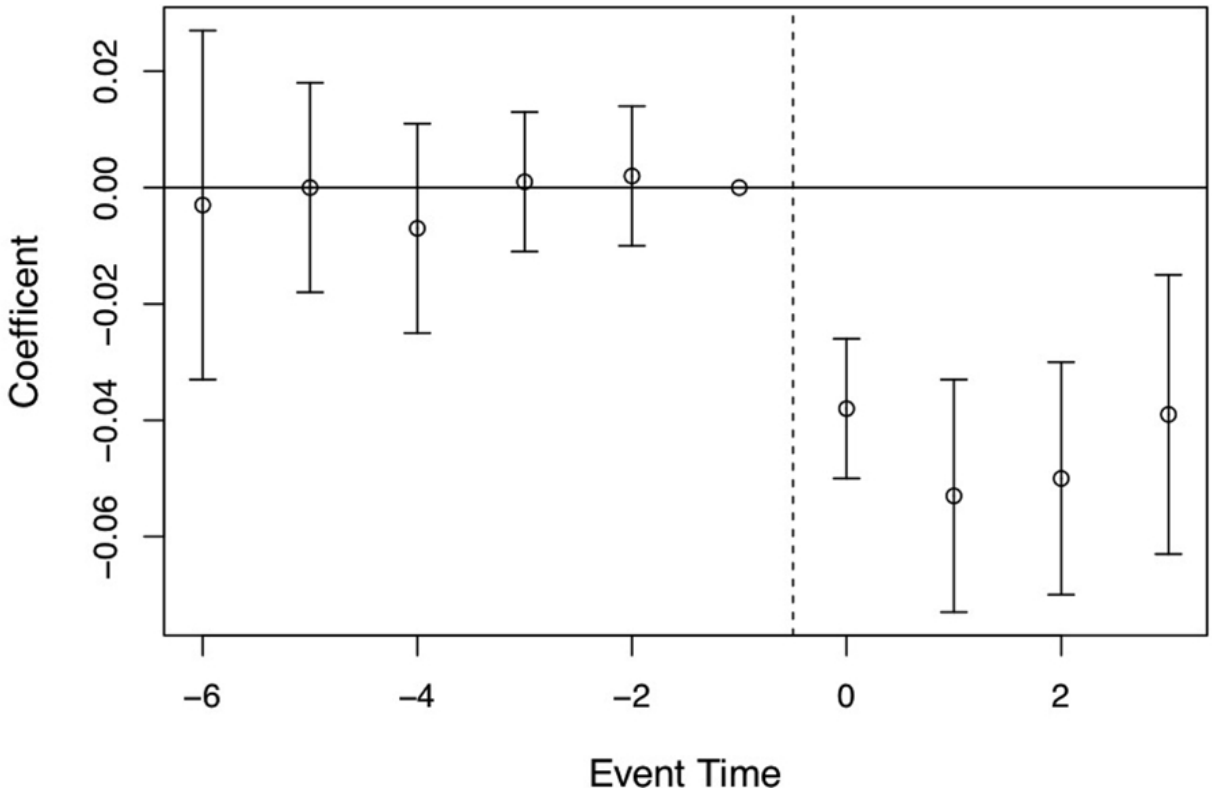


Figure 59. Estimates of Medicaid expansion's effects on the uninsured state using leads and lags in an event-study model. Miller, S., Altekruse, S., Johnson, N., and Wherry, L. R. (2019). Medicaid and mortality: New evidence from linked survey and administrative data. Working Paper No. 6081, National Bureau of Economic Research, Cambridge, MA. Reprinted with permission from authors.

But a picture like this is only as important as the thing that it is studying, and it is worth summarizing what Miller et al. [2019] have revealed here. The expansion of ACA Medicaid led to large swaths of people becoming eligible for Medicaid. In turn, they enrolled in Medicaid, which caused the uninsured rate to drop considerably. The authors then find amazingly using linked administrative data on death records that the expansion of ACA Medicaid led to a 0.13 percentage point decline in annual mortality, which is a 9.3 percent reduction over the mean. They go on to try and understand the mechanism (another key feature of this high-quality study) by which such amazing effects may have occurred, and conclude that Medicaid caused near-elderly individuals to receive treatment for life-threatening illnesses. I suspect we will be hearing about this study for many years.

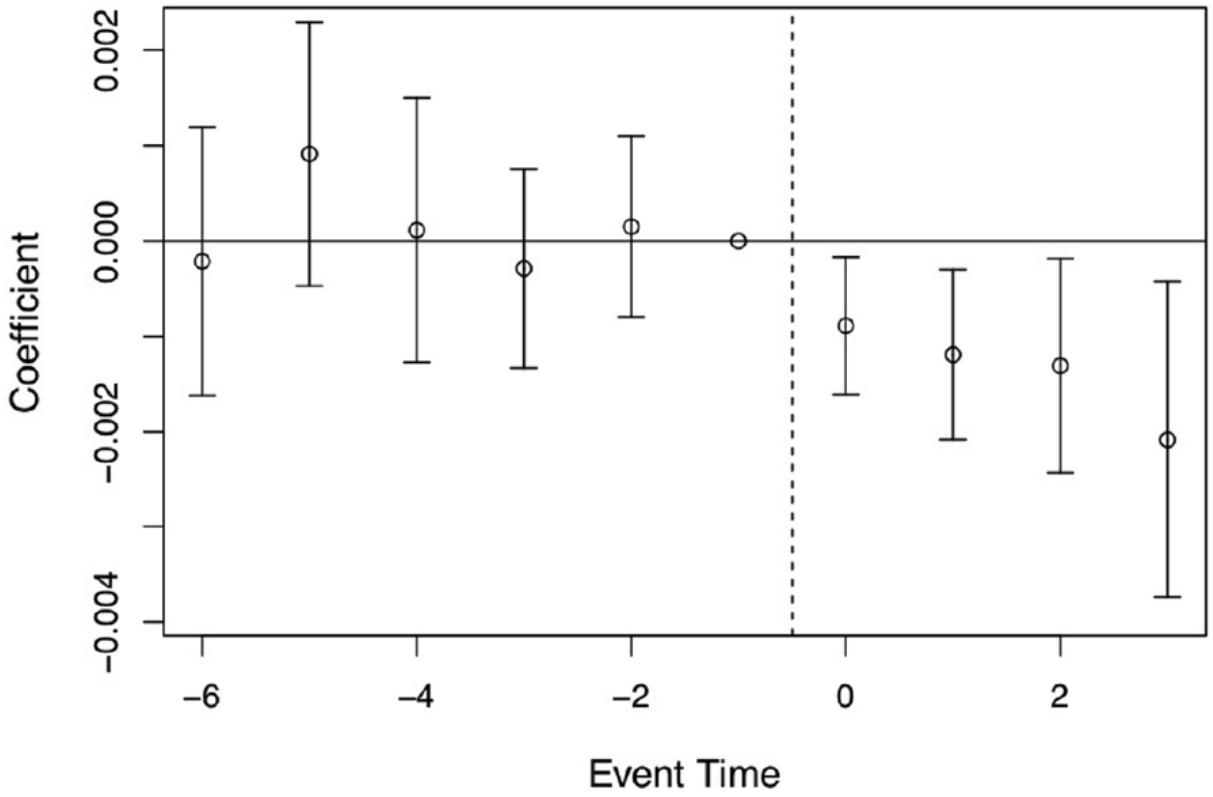


Figure 60. Estimates of Medicaid expansion's effects on on annual mortality using leads and lags in an event study model. Miller, S., Altekruse, S., Johnson, N., and Wherry, L. R. (2019). Medicaid and mortality: New evidence from linked survey and administrative data. Working Paper No. 6081, National Bureau of Economic Research, Cambridge, MA. Reprinted with permission from authors.

The Importance of Placebos in DD

There are several tests of the validity of a DD strategy. I have already discussed one—comparability between treatment and control groups on observable pre-treatment dynamics. Next, I will discuss other credible ways to evaluate whether estimated causal effects are credible by emphasizing the use of placebo falsification.

The idea of placebo falsification is simple. Say that you are finding some negative effect of the minimum wage on low-wage employment. Is the hypothesis true if we find evidence in favor? Maybe, maybe not. Maybe what would really help, though, is if you had in mind an alternative hypothesis and then tried to test that alternative hypothesis. If you cannot reject the null on the alternative

hypothesis, then it provides some credibility to your original analysis. For instance, maybe you are picking up something spurious, like cyclical factors or other unobservables not easily captured by a time or state fixed effects. So what can you do?

One candidate placebo falsification might simply be to use data for an alternative type of worker whose wages would not be affected by the binding minimum wage. For instance, minimum wages affect employment and earnings of low-wage workers as these are the workers who literally are hired based on the market wage. Without some serious general equilibrium gymnastics, the minimum wage should not affect the employment of higher wage workers, because the minimum wage is not binding on high wage workers. Since high- and low-wage workers are employed in very different sectors, they are unlikely to be substitutes. This reasoning might lead us to consider the possibility that higher wage workers *might* function as a placebo.

There are two ways you can go about incorporating this idea into our analysis. Many people like to be straightforward and simply fit the same DD design using high wage employment as the outcome. If the coefficient on minimum wages is zero when using high wage worker employment as the outcome, but the coefficient on minimum wages for low wage workers is negative, then we have provided stronger evidence that complements the earlier analysis we did when on the low wage workers. But there is another method that uses the within-state placebo for identification called the difference-in-differences-in-differences ("triple differences"). I will discuss that design now.

Triple differences. In our earlier analysis, we assumed that the only thing that happened to New Jersey after it passed the minimum wage was a common shock, T, but what if there were state-specific time shocks such as NJ_t or PA_t? Then even DD cannot recover the treatment effect. Let's see for ourselves using a modification of the simple minimum-wage table from earlier, which will include the

within-state workers who hypothetically were untreated by the minimum wage—the "high-wage workers."

Before the minimum-wage increase, low- and high-wage employment in New Jersey is determined by a group-specific New Jersey fixed effect (e.g., NJ_h). The same is true for Pennsylvania. But after the minimum-wage hike, four things change in New Jersey: national trends cause employment to change by T; New Jersey-specific time shocks change employment by NJ_t; generic trends in low-wage workers change employment by l_t; and the minimum-wage has some unknown effect D. We have the same setup in Pennsylvania except there is no minimum wage, and Pennsylvania experiences its own time shocks.

Table 74. Triple differences design.

States	Group	Period	Outcomes	D_1	D_2	D_3
NJ	Low-wage workers	After	$NJ_l + T + NJ_t + l_t + D$	$T + NJ_t + l_t + D$	$(l_t - h_t) + D$	D
		Before	NJ_l			
	High-wage workers	After	$NJ_h + T + NJ_t + h_t$	$T + NJ_t + h_t$		
		Before	NJ_h			
PA	Low-wage workers	After	$PA_l + T + PA_t + l_t$	$T + PA_t + l_t$	$l_t - h_t$	
		Before	PA_l			
	High-wage workers	After	$PA_h + T + PA_t + h_t$	$T + PA_t + h_t$		
		Before	PA_h			

Now if we take first differences for each set of states, we only eliminate the state fixed effect. The first difference estimate for New Jersey includes the minimum-wage effect, D, but is also hopelessly contaminated by confounders (i.e., $T + NJ_t + l_t$). So we take a second difference for each state, and doing so, we eliminate two of the confounders: T disappears and NJ_t disappears. But while this DD strategy has eliminated several confounders, it has also introduced new ones (i.e., $(l_t - h_t)$). This is the final source of selection bias that triple differences are designed to resolve. But, by differencing

Pennsylvania's second difference from New Jersey, the $(l_t - h_t)$ is deleted and the minimum-wage effect is isolated.

Now, this solution is not without its own set of unique parallel-trends assumptions. But one of the parallel trends here I'd like you to see is the $l_t - h_t$ term. This parallel trends assumption states that the effect can be isolated if the gap between high- and low-wage employment would've evolved similarly in the treatment state counterfactual as it did in the historical control states. And we should probably provide some credible evidence that this is true with leads and lags in an event study as before.

State-mandated maternity benefits. The triple differences design was first introduced by Gruber [1994] in a study of state-level policies providing maternity benefits. I present his main results in Table 75. Notice that he uses as his treatment group married women of childbearing age in treatment and control states, but he also uses a set of placebo units (older women and single men 20–40) as within-state controls. He then goes through the differences in means to get the difference-in-differences for each set of groups, after which he calculates the DDD as the difference between these two difference-in-differences.

Table 75. DDD Estimates of the Impact of State Mandates on Hourly Wages.

Location/year	Pre-law	Post-law	Difference
A. Treatment: Married women, 20–40yo			
Experimental states	1.547	1.513	−0.034
	(0.012)	(0.012)	(0.017)
Control states	1.369	1.397	0.028
	(0.010)	(0.010)	(0.014)
Difference	0.178	0.116	
	(0.016)	(0.015)	
Difference-in-difference		−0.062	
		(0.022)	
B. Control: Over 40 and Single Males 20–40			
Experimental states	1.759	1.748	−0.011
	(0.007)	(0.007)	(0.010)
Control states	1.630	1.627	−0.003
	(0.007)	(0.007)	(0.010)
Difference	1.09	1.21	
	(0.010)	(0.010)	
Difference-in-difference		−0.008	
		(0.014)	
DDD		−0.054	
		(0.026)	

Note: Standard errors in parentheses.

Ideally when you do a DDD estimate, the causal effect estimate will come from changes in the treatment units, not changes in the control units. That's precisely what we see in Gruber [1994]: the action comes from changes in the married women age 20–40 (−0.062); there's little movement among the placebo units (−0.008). Thus when we calculate the DDD, we know that most of that calculation is coming from the first DD, and not so much from the second. We emphasize this because DDD is really just another falsification exercise, and just as we would expect no effect had we done the DD on this placebo group, we hope that our DDD estimate is also based on negligible effects among the control group.

What we have done up to now is show how to use sample analogs and simple differences in means to estimate the treatment effect using DDD. But we can also use regression to control for additional covariates that perhaps are necessary to close backdoor paths and so forth. What does that regression equation look like? Both the regression itself, and the data structure upon which the regression is based, are complicated because of the stacking of different groups and the sheer number of interactions involved. Estimating a DDD model requires estimating the following regression:

$$Y_{ijt} = \alpha + \psi X_{ijt} + \beta_1 \tau_t + \beta_2 \delta_j + \beta_3 D_i + \beta_4 (\delta \times \tau)_{jt} + \beta_5 (\tau \times D)_{ti} + \beta_6 (\delta \times D)_{ij} + \beta_7 (\delta \times \tau \times D)_{ijt} + \varepsilon_{ijt}$$

where the parameter of interest is β_7. First, notice the additional subscript, *j*. This *j* indexes whether it's the main category of interest (e.g., low-wage employment) or the within-state comparison group (e.g., high-wage employment). This requires a stacking of the data into a panel structure by group, as well as state. Second, the DDD model requires that you include all possible interactions across the group dummy δ_j, the post-treatment dummy τ_t and the treatment state dummy D_i. The regression must include each dummy independently, each individual interaction, and the triple differences interaction. One of these will be dropped due to multicollinearity, but I include them in the equation so that you can visualize all the factors used in the product of these terms.

Abortion legalization and long-term gonorrhea incidence. Now that we know a little about the DD design, it would probably be beneficial to replicate a paper. And since the DDD requires reshaping panel data multiple times, that makes working through a detailed replication even more important. The study we will be replicating is Cunningham and Cornwell [2013], one of my first publications and the third chapter of my dissertation. Buckle up, as this will be a bit of a roller-coaster ride.

Gruber et al. [1999] was the beginning of what would become a controversial literature in reproductive health. They wanted to know the characteristics of the marginal child aborted had that child reached their teen years. The authors found that the marginal counterfactual child aborted was 60% more likely to grow up in a single-parent household, 50% more likely to live in poverty, and 45% more likely to be a welfare recipient. Clearly there were strong selection effects related to early abortion whereby it selected on families with fewer resources.

Their finding about the marginal child led John Donohue and Steven Levitt to wonder if there might be far-reaching effects of abortion legalization given the strong selection associated with its usage in the early 1970s. In Donohue and Levitt [2001], the authors argued that they had found evidence that abortion legalization had also led to massive declines in crime rates. Their interpretation of the results was that abortion legalization had reduced crime by removing high-risk individuals from a birth cohort, and as that cohort aged, those counterfactual crimes disappeared. Levitt [2004] attributed as much as 10% of the decline in crime between 1991 and 2001 to abortion legalization in the 1970s.

This study was, not surprisingly, incredibly controversial—some of it warranted but some unwarranted. For instance, some attacked the paper on ethical grounds and argued the paper was revitalizing the pseudoscience of eugenics. But Levitt was careful to focus only on the scientific issues and causal effects and did not offer policy advice based on his own private views, whatever those may be.

But some of the criticism the authors received was legitimate precisely because it centered on the research design and execution itself. Joyce [2004], Joyce [2009], and Foote and Goetz [2008] disputed the abortion-crime findings—some through replication exercises using different data, some with different research designs, and some through the discovery of key coding errors and erroneous variable construction.

One study in particular challenged the whole enterprise of estimating longrun improvements due to abortion legalization. For instance, Ted Joyce, an expert on reproductive health, cast doubt on

the abortion-crime hypothesis using a DDD design [Joyce, 2009]. In addition to challenging Donohue and Levitt [2001], Joyce also threw down a gauntlet. He argued that if abortion legalization had such extreme negative selection as claimed by by Gruber et al. [1999] and Donohue and Levitt [2001], then it shouldn't show up just in crime. It should show up *everywhere*. Joyce writes:

> If abortion lowers homicide rates by 20–30%, then it is likely to have affected an entire spectrum of outcomes associated with well-being: infant health, child development, schooling, earnings and marital status. Similarly, the policy implications are broader than abortion. Other interventions that affect fertility control and that lead to fewer unwanted births—contraception or sexual abstinence—have huge potential payoffs. In short, a causal relationship between legalized abortion and crime has such significant ramifications for social policy and at the same time is so controversial, that further assessment of the identifying assumptions and their robustness to alternative strategies is warranted. [112]

Cunningham and Cornwell [2013] took up Joyce's challenge. Our study estimated the effects of abortion legalization on long-term gonorrhea incidence. Why gonorrhea? For one, single-parent households are a risk factor that lead to earlier sexual activity and unprotected sex, and Levine et al. [1999] found that abortion legalization caused teen childbearing to fall by 12%. Other risky outcomes had been found by numerous authors. Charles and Stephens [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.

My research design differed from Donohue and Levitt [2001] in that they used state-level lagged values of an abortion ratio, whereas I used difference-in-differences. My design exploited the early repeal of abortion in five states in 1970 and compared those states to the states that were legalized under *Roe v. Wade* in 1973. To do this, I needed cohort-specific data on gonorrhea incidence by state and year, but as those data are not collected by the CDC, I had to settle for second best. That second best was the CDC's gonorrhea data broken into five-year age categories (e.g., age 15–19, age 20–24).

But this might still be useful because even with aggregate data, it might be possible to test the model I had in mind.

To understand this next part, which I consider the best part of my study, you must first accept a basic view of science that good theories make very specific falsifiable hypotheses. The more specific the hypothesis, the more convincing the theory, because if we find evidence exactly where the theory predicts, a Bayesian is likely to update her beliefs towards accepting the theory's credibility. Let me illustrate what I mean with a brief detour involving Albert Einstein's theory of relativity.

Einstein's theory made several falsifiable hypotheses. One of them involved a precise prediction of the warping of light as it moved past a large object, such as a star. The problem was that testing this theory involved observing distance between stars at night and comparing it to measurements made during the day as the starlight moved past the sun. Problem was, the sun is too bright in the daytime to see the stars, so those critical measurements can't be made. But Andrew Crommelin and Arthur Eddington realized the measurements could be made using an ingenious natural experiment. That natural experiment was an eclipse. They shipped telescopes to different parts of the world under the eclipse's path so that they had multiple chances to make the measurements. They decided to measure the distances of a large cluster of stars passing by the sun when it was dark and then immediately during an eclipse (Figure 61). That test was over a decade after Einstein's work was first published [Coles, 2019]. Think about it for a second—Einstein's theory by deduction is making predictions about phenomena that no one had ever really observed before. If this phenomena turned out to exist, then how couldn't the Bayesian update her beliefs and accept that the theory was credible? Incredibly, Einstein was right—just as he predicted, the apparent position of these stars shifted when moving around the sun. Incredible!

So what does that have to do with my study of abortion legalization and gonorrhea? The theory of abortion legalization having strong selection effects on cohorts makes very specific predictions about the shape of observed treatment effects. And if we found evidence

for that shape, we'd be forced to take the theory seriously. So what what were these unusual yet testable predictions exactly?

The testable prediction from the staggered adoption of abortion legalization concerned the age-year-state profile of gonorrhea. The early repeal of abortion by five states three years before the rest of the country predicts lower incidence among 15- to 19-year-olds in the repeal states only during the 1986–1992 period relative to their *Roe* counterparts as the treated cohorts aged. That's not really all that special a prediction though. Maybe something happens in those same states fifteen to nineteen years later that isn't controlled for, for instance. What else?

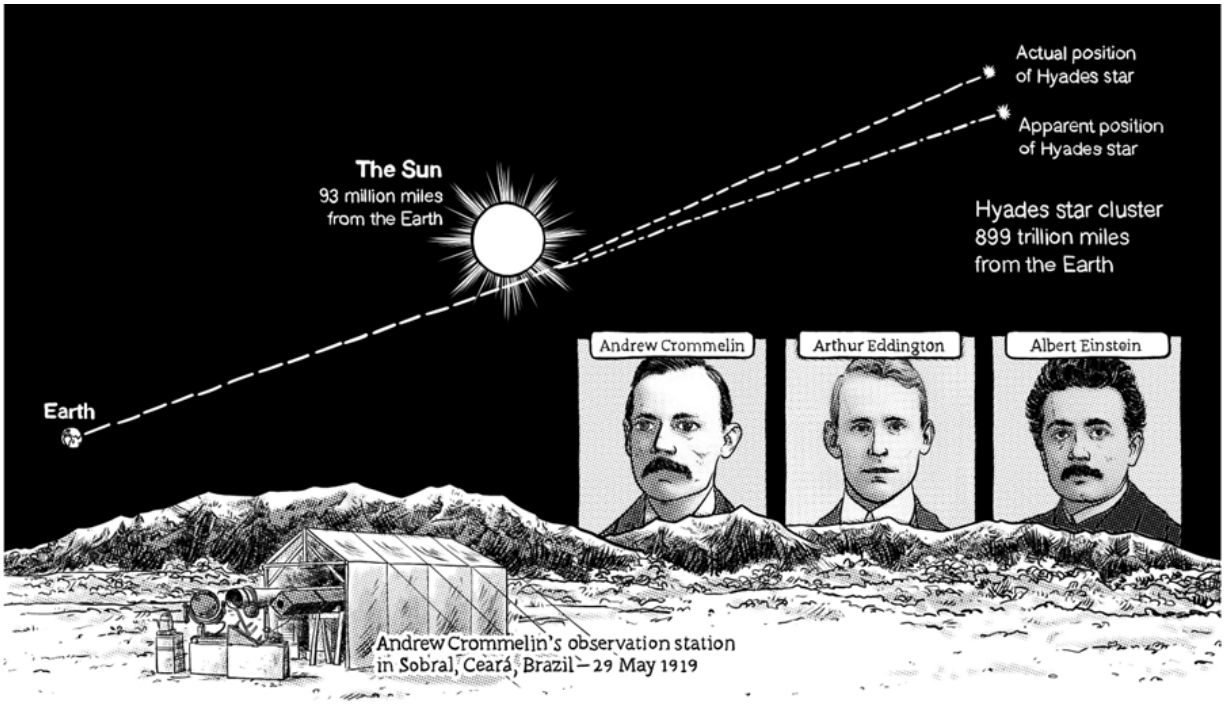


Figure 61. Light bending around the sun, predicted by Einstein, and confirmed in a natural experiment involving an eclipse. Artwork by Seth Hahne ©2020.

The abortion legalization theory also predicted the *shape* of the observed treatment effects in this particular staggered adoption. Specifically, we should observe nonlinear treatment effects. These treatment effects should be increasingly negative from 1986 to 1989, plateau from 1989 to 1991, then gradually dissipate until 1992. In

other words, the abortion legalization hypothesis predicts a parabolic treatment effect as treated cohorts move through the age distribution. All coefficients on the DD coefficients beyond 1992 should be zero and/or statistically insignificant.

I illustrate these predictions in Figure 62. The top horizontal axis shows the year of the panel, the vertical axis shows the age in calendar years, and the cells show the cohort for a given person of a certain age in that given year. For instance, consider a 15-year-old in 1985. She was born in 1970. A 15-year-old in 1986 was born in 1971. A 15-year-old in 1987 was born in 1972, and so forth. I mark the cohorts who were treated by either repeal or *Roe* in different shades of gray.

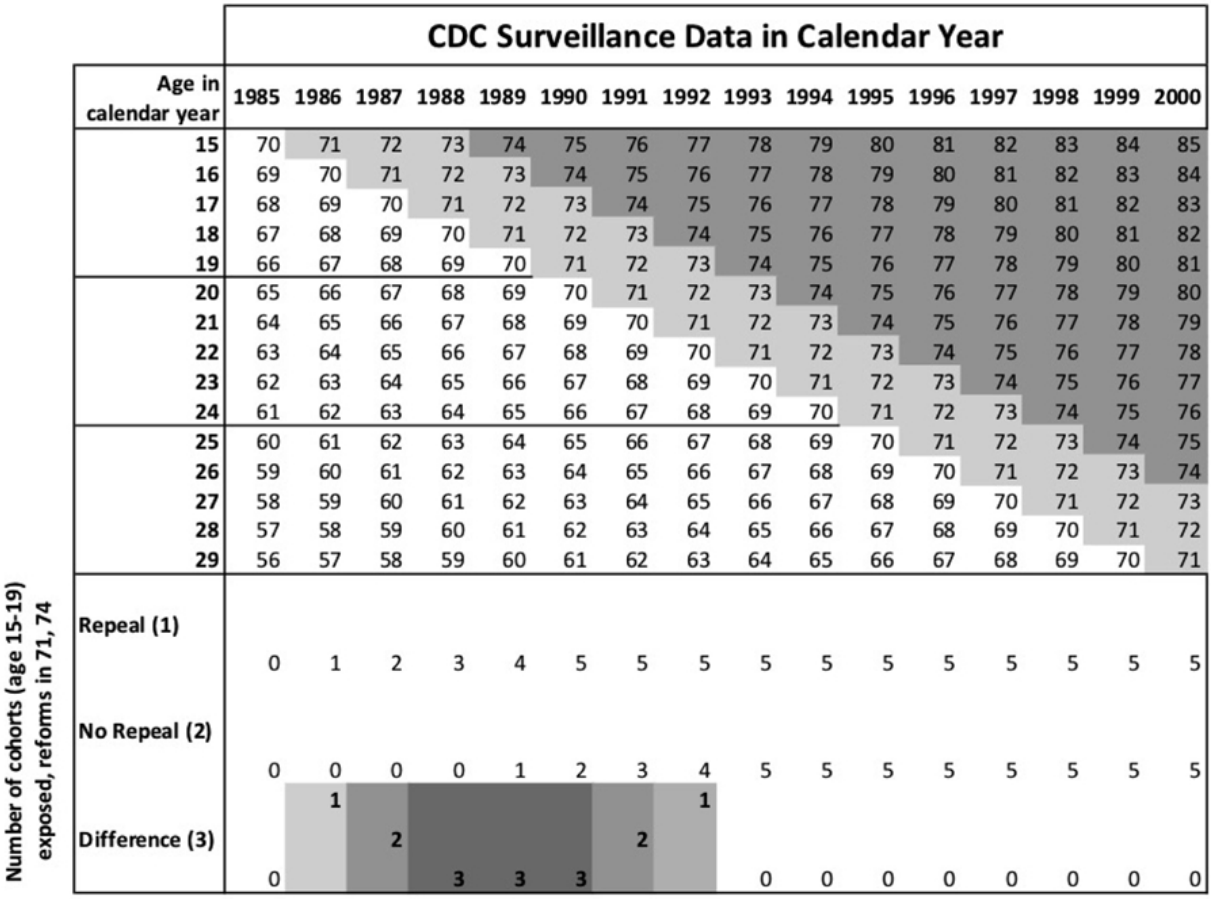


Figure 62. Theoretical predictions of abortion legalization on age profiles of gonorrhea incidence. Reprinted from Cunningham, S. and Cornwell, C. (2013). "The Long-Run Effect of Abortion on Sexually Transmitted Infections." *American Law and Economics Review*, 15(1):381–407. Permission from Oxford University Press.

The theoretical predictions of the staggered rollout is shown at the bottom of Figure 62. In 1986, only one cohort (the 1971 cohort) was treated and only in the repeal states. Therefore, we should see small declines in gonorrhea incidence among 15-year-olds in 1986 relative to *Roe* states. In 1987, two cohorts in our data are treated in the repeal states relative to *Roe*, so we should see larger effects in absolute value than we saw in 1986. But from 1988 to 1991, we should at most see only three *net* treated cohorts in the repeal states because starting in 1988, the *Roe* state cohorts enter and begin erasing those differences. Starting in 1992, the effects should get

smaller in absolute value until 1992, beyond which there should be no difference between repeal and *Roe* states.

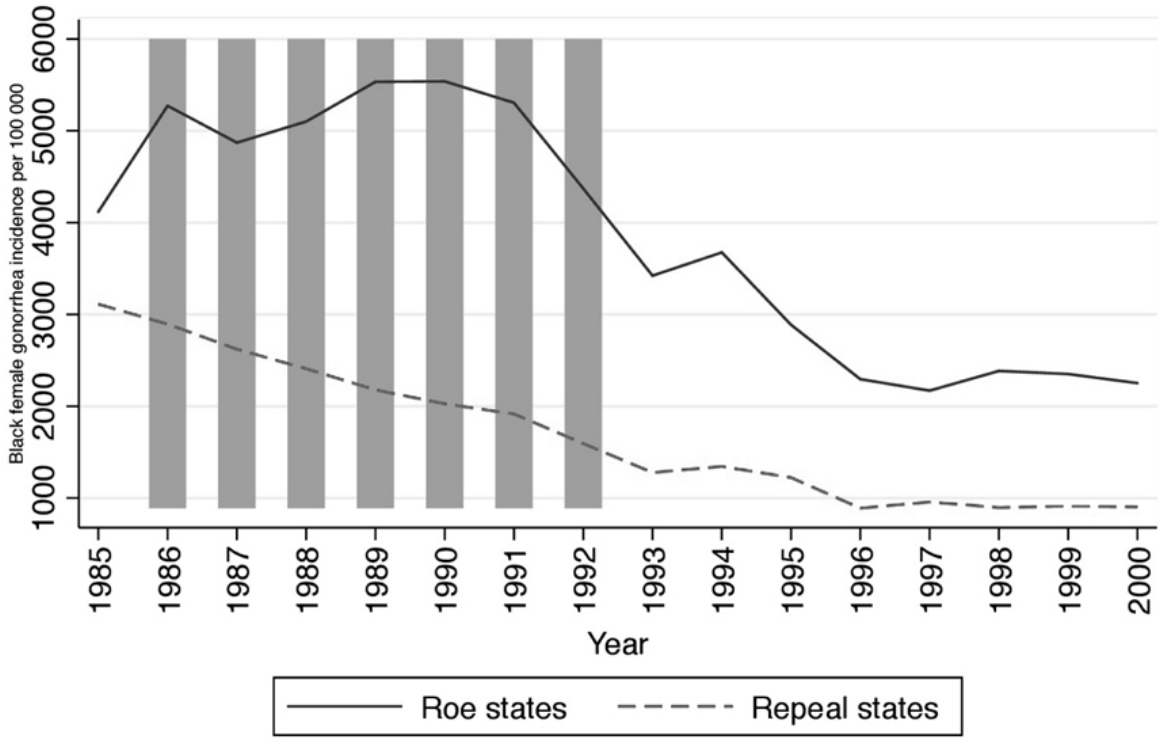


Figure 63. Differences in gonorrhea incidence among black females between repeal and *Roe* cohorts expressed as coefficient plots. Reprinted from Cunningham, S. and Cornwell, C. (2013). "The Long-Run Effect of Abortion on Sexually Transmitted Infections." *American Law and Economics Review*, 15(1):381–407. Permission from Oxford University Press.

It is interesting that something so simple as a staggered policy rollout should provide two testable hypotheses that together can provide some insight into whether there is credibility to the negative selection in abortion legalization story. If we cannot find evidence for a negative parabola during this specific, narrow window, then the abortion legalization hypothesis has one more nail in its coffin.

A simple graphic for gonorrhea incidence among black 15- to 19-year-olds can help illustrate our findings. Remember, a picture is worth 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. So notice what the raw data looks like in Figure 63.

First let's look at the raw data. I have shaded the years corresponding to the window where we expect to find effects. In Figure 63, we see the dynamics that will ultimately be picked up in the regression coefficients—the *Roe* states experienced a large and sustained gonorrhea epidemic that only waned once the treated cohorts emerged and overtook the entire data series.

Now let's look at regression coefficients. Our estimating equation is as follows:

$$Y_{st} = \beta_1 Repeals + \beta_2 DT_t + \beta_{3t} Repeal_s \times DT_t + X_{st}\psi + \alpha_s DS_s + \varepsilon_{st}$$

where Y is the log number of new gonorrhea cases for 15- to 19-year-olds (per 100,000 of the population); $Repeal_s$ equals 1 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. In the paper, I sometimes included state-specific linear trends, but for this analysis, I present the simpler model. Finally, ε_{st} is a structural error term assumed to be conditionally independent of the regressors. All standard errors, furthermore, were clustered at the state level allowing for arbitrary serial correlation.

I present the plotted coefficients from this regression for simplicity (and because pictures can be so powerful) in Figure 64. As can be seen in Figure 64, there is a negative effect during the window where *Roe* has not *fully* caught up, and that negative effect forms a parabola—just as our theory predicted.

Now, a lot of people might be done, but if you are reading this book, you have revealed that you are not like a lot of people. *Credibly* identified causal effects requires both finding effects, and ruling out alternative explanations. This is necessary because the fundamental problem of causal inference keeps us blind to the truth. But one way to alleviate some of that doubt is through rigorous placebo analysis. Here I present evidence from a triple difference in which an untreated cohort is used as a within-state control.

We chose the 25- to 29-year-olds in the same states as within-state comparison groups instead of 20- to 24-year-olds after a lot of

thought. Our reasoning was that we needed an age group that was close enough to capture common trends but far enough so as not to violate SUTVA. Since 15- to 19-year-olds were more likely than 25-to 29-year-olds to have sex with 20- to 24-year-olds, we chose the slightly older group as the within-stage control. But there's a trade-off here. Choose a group too close and you get SUTVA violations. Choose a group too far and they no longer can credibly soak up the heterogeneities you're worried about. The estimating equation for this regression is:

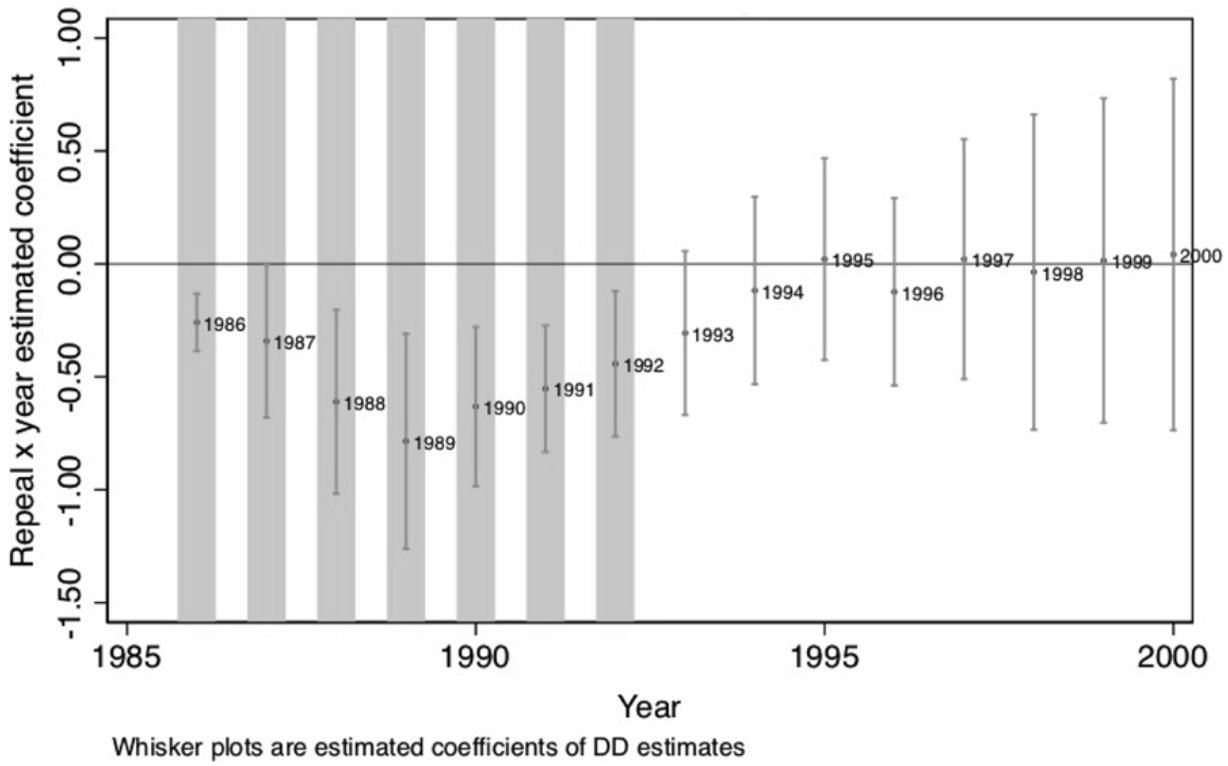


Figure 64. Coefficients and standard errors from DD regression equation.

$$Y_{ast} = \beta_1 \text{ Repeal}_s + \beta_2 DT_t + \beta_{3t} \text{ Repeal}_s \cdot DT_t + \delta_1 DA + \delta_2 \text{ Repeal}_s \cdot DA + \delta_{3t} DA \cdot DT_t + \delta_{4t} \text{ Repeal}_s \cdot DA \cdot DT_t + X_{st}\xi + \alpha_{1s} DS_s + \alpha_{2s} DS_s \cdot DA + \gamma_1 t + \gamma_{2s} DS_s \cdot t + \gamma_3 DA \cdot t + \gamma_{4s} DS_s \cdot DA \cdot t + \epsilon_{ast},$$

where the DDD parameter we are estimating is δ_{4t}—the full interaction. In case this wasn't obvious, there are 7 separate dummies because our DDD parameter has all three interactions.

Thus since there are eight combinations, we had to drop one as the omitted group, and control separately for the other seven. Here we present the table of coefficients. Note that the effect should be concentrated only among the treatment years as before, and second, it should form a parabola. The results are presented in Figure 65.

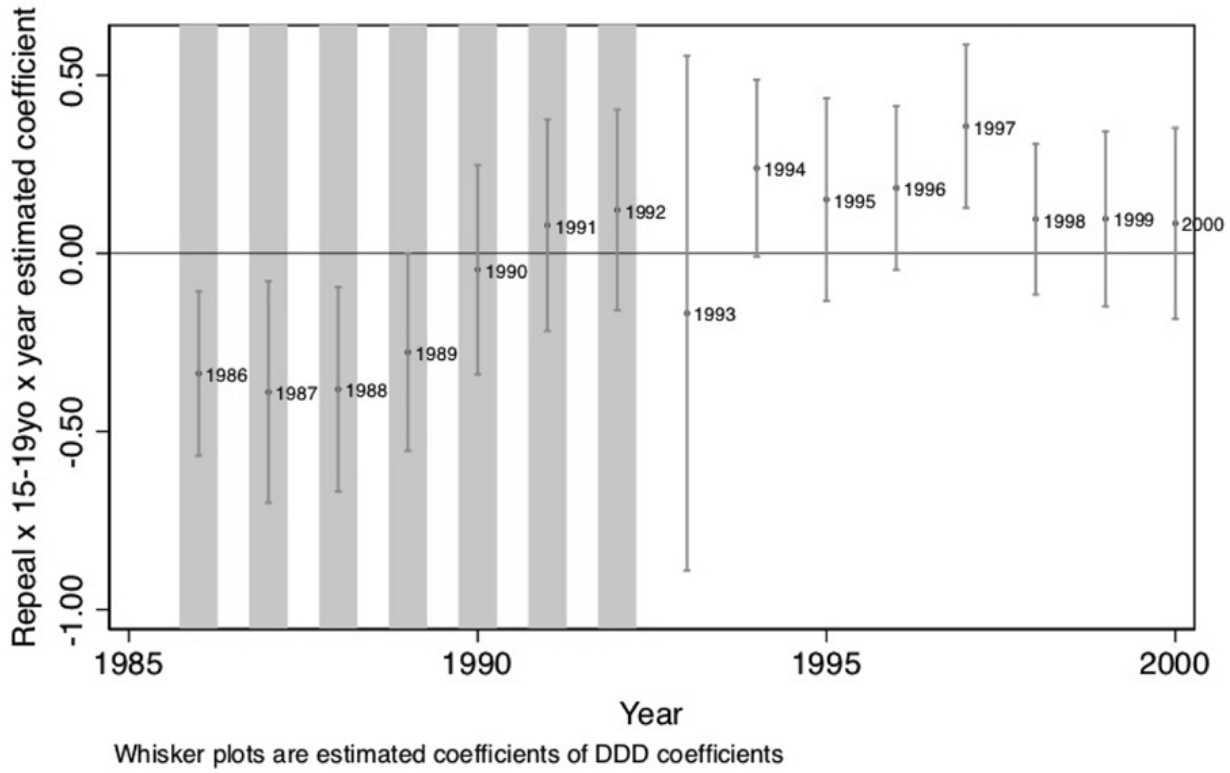


Figure 65. DDD coefficients of abortion legalization on 15- to 19-year-old Black female log gonorrhea rates.

Here we see the prediction start to break down. Though there are negative effects for years 1986 to 1990, the 1991 and 1992 coefficients are positive, which is not consistent with our hypothesis. Furthermore, only the first four coefficients are statistically significant. Nevertheless, given the demanding nature of DDD, perhaps this is a small victory in favor of Gruber et al. [1999] and Donohue and Levitt [2001]. Perhaps the theory that abortion legalization had strong selection effects on cohorts has some validity.

Putting aside whether you believe the results, it is still valuable to replicate the results based on this staggered design. Recall that I said the DDD design requires stacking the data, which may seem like a bit of a black box, so I'd like to examine these data now.[7]

STATA

abortion_dd.do

```
* DD estimate of 15-19 year olds in repeal states vs Roe states
use https://github.com/scunning1975/mixtape/raw/master/abortion.dta, clear
xi: reg lnr i.repeal*i.year i.fip acc ir pi alcohol crack poverty income ur if bf15==1
    ↪ [aweight=totpop], cluster(fip)

* ssc install parmest, replace

parmest, label for(estimate min95 max95 %8.2f) li(parm label estimate min95
    ↪ max95) saving(bf15_DD.dta, replace)

use ./bf15_DD.dta, replace

keep in 17/31

gen             year=1986 in 1
replace year=1987 in 2
replace year=1988 in 3
replace year=1989 in 4
replace year=1990 in 5
replace year=1991 in 6
replace year=1992 in 7
replace year=1993 in 8
replace year=1994 in 9
replace year=1995 in 10
replace year=1996 in 11
replace year=1997 in 12
replace year=1998 in 13
replace year=1999 in 14
replace year=2000 in 15

sort year

twoway (scatter estimate year, mlabel(year) mlabsize(vsmall) msize(tiny)) (rcap
    ↪ min95 max95 year, msize(vsmall)), ytitle(Repeal x year estimated
    ↪ coefficient) yscale(titlegap(2)) yline(0, lwidth(vvvthin) lcolor(black))
    ↪ xtitle(Year) xline(1986 1987 1988 1989 1990 1991 1992, lwidth(vvvthick)
    ↪ lpattern(solid) lcolor(ltblue)) xscale(titlegap(2)) title(Estimated effect of
    ↪ abortion legalization on gonorrhea) subtitle(Black females 15-19 year-olds)
    ↪ note(Whisker plots are estimated coefficients of DD estimator from Column
    ↪ b of Table 2.) legend(off)
```

R

abortion_dd.R

```
#-- DD estimate of 15-19 year olds in repeal states vs Roe states
library(tidyverse)
library(haven)
library(estimatr)

read_data <- function(df)
{
 full_path <- paste("https://raw.github.com/scunning1975/mixtape/master/",
                    df, sep = "")
 df <- read_dta(full_path)
 return(df)
}

abortion <- read_data("abortion.dta") %>%
 mutate(
  repeal = as_factor(repeal),
  year   = as_factor(year),
  fip    = as_factor(fip),
  fa     = as_factor(fa),
 )

reg <- abortion %>%
 filter(bf15 == 1) %>%
 lm_robust(lnr ~ repeal*year + fip + acc + ir + pi + alcohol+ crack + poverty+
 ↪  income+ ur,
           data = ., weights = totpop, clusters = fip)

abortion_plot <- tibble(
 sd = reg$std.error[-1:-75],
 mean = reg$coefficients[-1:-75],
 year = c(1986:2000))

abortion_plot %>%
 ggplot(aes(x = year, y = mean)) +
 geom_rect(aes(xmin=1986, xmax=1992, ymin=-Inf, ymax=Inf), fill = "cyan", alpha
 ↪  = 0.01)+
 geom_point()+
 geom_text(aes(label = year), hjust=-0.002, vjust = -0.03)+
 geom_hline(yintercept = 0) +
 geom_errorbar(aes(ymin = mean - sd*1.96, ymax = mean + sd*1.96), width = 0.2,
               position = position_dodge(0.05))
```

The second line estimates the regression equation. The dynamic DD coefficients are captured by the repeal-year interactions. These are the coefficients we used to create box plots in Figure 64. You can check these yourself.

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 here we only focus on the black females aged 15–19 because the purpose of this section is to help you understand the estimation. I encourage you to play around with this model to see

how robust the effects are in your mind using only this linear estimation.

But now I want to show you the code for estimating a triple difference model. Some reshaping had to be done behind the scenes for this data structure, but it would take too long to post that here. For now, I will simply produce the commands that produce the black female result, and I encourage you to explore the panel data structure so as to familiarize yourself with the way in which the data are organized.

Notice that some of these were already interactions (e.g., yr), which was my way to compactly include all of the interactions. I did this primarily to give myself more control over what variables I was using. But I encourage you to study the data structure itself so that when you need to estimate your own DDD, you'll have a good handle on what form the data must be in in order to execute so many interactions.

STATA

abortion_ddd.do

```
1  use https://github.com/scunning1975/mixtape/raw/master/abortion.dta, clear
2
3  * DDD estimate for 15-19 year olds vs. 20-24 year olds in repeal vs Roe states
4  gen yr=(repeal) & (younger==1)
5  gen wm=(wht==1) & (male==1)
6  gen wf=(wht==1) & (male==0)
7  gen bm=(wht==0) & (male==1)
8  gen bf=(wht==0) & (male==0)
9  char year[omit] 1985
10 char repeal[omit] 0
11 char younger[omit] 0
```

(continued)

STATA *(continued)*

```
11  char fip[omit] 1
12  char fa[omit] 0
13  char yr[omit] 0
14  xi: reg 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 bf==1 & (age==15 | age==25)
    ↪ [aweight=totpop], cluster(fip)
15
16  parmest, label for(estimate min95 max95 %8.2f) li(parm label estimate min95
    ↪ max95) saving(bf15_DDD.dta, replace)
17
18  use ./bf15_DDD.dta, replace
19
20  keep in 82/96
21
22  gen             year=1986 in 1
23  replace year=1987 in 2
24  replace year=1988 in 3
25  replace year=1989 in 4
26  replace year=1990 in 5
27  replace year=1991 in 6
28  replace year=1992 in 7
29  replace year=1993 in 8
30  replace year=1994 in 9
31  replace year=1995 in 10
32  replace year=1996 in 11
33  replace year=1997 in 12
34  replace year=1998 in 13
35  replace year=1999 in 14
36  replace year=2000 in 15
37
38  sort year
39
40  twoway (scatter estimate year, mlabel(year) mlabsize(vsmall) msize(tiny)) (rcap
    ↪ min95 max95 year, msize(vsmall)), ytitle(Repeal x 20-24yo x year estimated
    ↪ coefficient) yscale(titlegap(2)) yline(0, lwidth(vvvthin) lcolor(black))
    ↪ xtitle(Year) xline(1986 1987 1988 1989 1990 1991 1992, lwidth(vvvthick)
    ↪ lpattern(solid) lcolor(ltblue)) xscale(titlegap(2)) title(Estimated effect of
    ↪ abortion legalization on gonorrhea) subtitle(Black females 15-19 year-olds)
    ↪ note(Whisker plots are estimated coefficients of DDD estimator from
    ↪ Column b of Table 2.) legend(off)
41
```

R

abortion_ddd.R

```r
library(tidyverse)
library(haven)
library(estimatr)

read_data <- function(df)
{
  full_path <- paste("https://raw.github.com/scunning1975/mixtape/master/",
                     df, sep = "")
  df <- read_dta(full_path)
  return(df)
}

abortion <- read_data("abortion.dta") %>%
  mutate(
    repeal  = as_factor(repeal),
    year    = as_factor(year),
    fip     = as_factor(fip),
    fa      = as_factor(fa),
    younger = as_factor(younger),
    yr      = as_factor(case_when(repeal == 1 & younger == 1 ~ 1, TRUE ~ 0)),
    wm      = as_factor(case_when(wht == 1 & male == 1 ~ 1, TRUE ~ 0)),
    wf      = as_factor(case_when(wht == 1 & male == 0 ~ 1, TRUE ~ 0)),
    bm      = as_factor(case_when(wht == 0 & male == 1 ~ 1, TRUE ~ 0)),
    bf      = as_factor(case_when(wht == 0 & male == 0 ~ 1, TRUE ~ 0))
  ) %>%
  filter(bf == 1 & (age == 15 | age == 25))

regddd <- lm_robust(lnr ~ repeal*year + younger*repeal + younger*year + yr*year
  ↪  + fip*t + acc + ir + pi + alcohol + crack + poverty + income + ur,
                    data = abortion, weights = totpop, clusters = fip)

abortion_plot <- tibble(
  sd = regddd$std.error[110:124],
  mean = regddd$coefficients[110:124],
  year = c(1986:2000))

```

(continued)

R *(continued)*

```r
abortion_plot %>%
  ggplot(aes(x = year, y = mean)) +
  geom_rect(aes(xmin=1986, xmax=1992, ymin=-Inf, ymax=Inf), fill = "cyan", alpha
  ↪  = 0.01)+
  geom_point()+
  geom_text(aes(label = year), hjust=-0.002, vjust = -0.03)+
  geom_hline(yintercept = 0) +
  geom_errorbar(aes(ymin = mean-sd*1.96, ymax = mean+sd*1.96), width = 0.2,
                position = position_dodge(0.05))
```

Going beyond Cunningham and Cornwell [2013]. The US experience with abortion legalization predicted a parabola from 1986 to 1992 for 15- to 19-year-olds, and using a DD design, that's what I found. I also estimated the effect using a DDD design, and while the effects weren't as pretty as what I found with DD, there appeared to be something going on in the general vicinity of where the model predicted. So boom goes the dynamite, right? Can't we be done finally? Not quite.

Whereas my original study stopped there, I would like to go a little farther. The reason can be seen in the following Figure 66. This is a modified version of Figure 62, with the main difference being I have created a new parabola for the 20- to 24-year-olds.

Look carefully at Figure 66. Insofar as the early 1970s cohorts were treated in utero with abortion legalization, then we should see not just a parabola for the 15- to 19-year-olds for 1986 to 1992 but also for the 20- to 24-year-olds for years 1991 to 1997 as the cohorts continued to age.[8]

I did not examine the 20- to 24-year-old cohort when I first wrote this paper because at that time I doubted that the selection effects for risky sex would persist into adulthood given that youth display considerable risk-taking behavior. But with time come new perspectives, and these days I don't have strong priors that the selection effects would necessarily vanish after teenage years. So I'd like to conduct that additional analysis here and now for the first time. Let's estimate the same DD model as before, only for Black females aged 20–24.

STATA

abortion_dd2.do

```
use https://github.com/scunning1975/mixtape/raw/master/abortion.dta, clear

* Second DD model for 20-24 year old black females
char year[omit] 1985
xi: reg lnr i.repeal*i.year i.fip acc ir pi alcohol crack poverty income ur if (race==2
    ↪  & sex==2 & age==20) [aweight=totpop], cluster(fip)
```

R

abortion_dd2.R

```
library(tidyverse)
library(haven)
library(estimatr)

read_data <- function(df)
{
  full_path <- paste("https://raw.github.com/scunning1975/mixtape/master/",
                     df, sep = "")
  df <- read_dta(full_path)
  return(df)
}

abortion <- read_data("abortion.dta") %>%
  mutate(
    repeal = as_factor(repeal),
    year   = as_factor(year),
    fip    = as_factor(fip),
    fa     = as_factor(fa),
  )

reg <- abortion %>%
  filter(race == 2 & sex == 2 & age == 20) %>%
  lm_robust(lnr ~ repeal*year + fip + acc + ir + pi + alcohol+ crack + poverty+
    ↪  income+ ur,
            data = ., weights = totpop, clusters = fip)
```

As before, we will focus just on the coefficient plots. We show that in Figure 67. There are a couple of things about this regression output that are troubling. First, there is a negative parabola showing up where there wasn't necessarily one predicted—the 1986–1992 period. Note that is the period where only the 15- to 19-year-olds were the treated cohorts, suggesting that our 15- to 19-year-old analysis was picking up something other than abortion legalization. But that was also the justification for using DDD, as clearly something else is going on in the repeal versus *Roe* states during

those years that we cannot adequately control for with our controls and fixed effects.

	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 20-24) exposed, reforms in 71, 74																
Repeal (1)	0	0	0	0	0	0	1	2	3	4	5	5	5	5	5	5
No Repeal (2)	0	0	0	0	0	0	0	0	0	1	2	3	4	5	5	5
Difference (3)	0	0	0	0	0	0	1	2	3	3	3	2	1	0	0	0

Figure 66. Theoretical predictions of abortion legalization on age profiles of gonorrhea incidence for 20–24-year-olds.

The second thing to notice is that there is *no* parabola in the treatment window for the treatment cohort. The effect sizes are negative in the beginning, but shrink in absolute value when they should be growing. In fact, the 1991 to 1997 period is one of convergence to zero, not divergence between these two sets of states.

But as before, maybe there are strong trending unobservables for all groups masking the abortion legalization effect. To check, let's use my DDD strategy with the 25- to 29-year-olds as the within-state control group. We can implement this by using the Stata code, abortion_ddd2.do and abortion_ddd2.R.

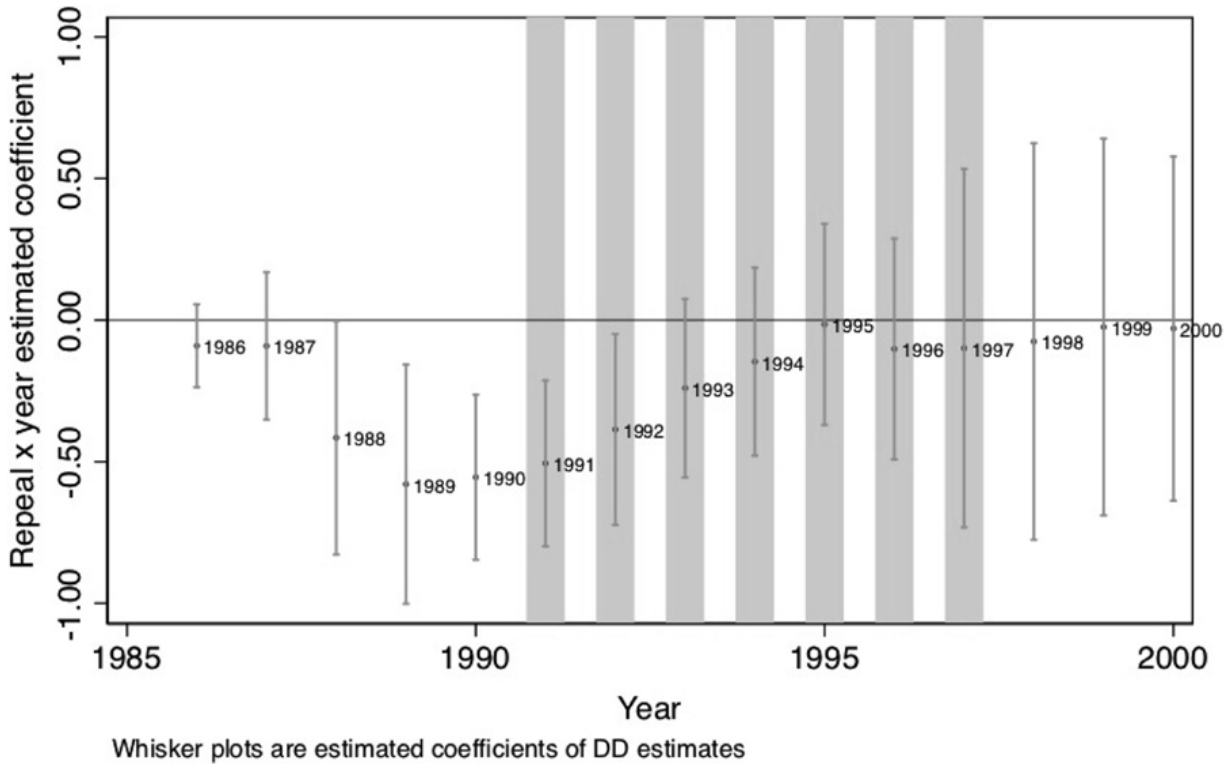


Figure 67. Coefficients and standard errors from DD regression equation for the 20- to 24-year-olds.

STATA

abortion_ddd2.do

```
1  use https://github.com/scunning1975/mixtape/raw/master/abortion.dta, clear
2
3  * Second DDD model for 20-24 year olds vs 25-29 year olds black females in
   ↪  repeal vs Roe states
4  gen younger2 = 0
5  replace younger2 = 1 if age == 20
6  gen yr2=(repeal==1) & (younger2==1)
7  gen wm=(wht==1) & (male==1)
8  gen wf=(wht==1) & (male==0)
9  gen bm=(wht==0) & (male==1)
10 gen bf=(wht==0) & (male==0)
11 char year[omit] 1985
12 char repeal[omit] 0
13 char younger2[omit] 0
14 char fip[omit] 1
15 char fa[omit] 0
16 char yr2[omit] 0
17 xi: reg lnr i.repeal*i.year i.younger2*i.repeal i.younger2*i.year i.yr2*i.year i.fip*t
   ↪  acc pi ir alcohol crack  poverty income ur if bf==1 & (age==20 | age==25)
   ↪  [aweight=totpop], cluster(fip)
```

R

abortion_ddd2.R

```
library(tidyverse)
library(haven)
library(estimatr)

read_data <- function(df)
{
  full_path <- paste("https://raw.github.com/scunning1975/mixtape/master/",
                     df, sep = "")
  df <- read_dta(full_path)
  return(df)
}

abortion <- read_data("abortion.dta") %>%
  mutate(
    repeal   = as_factor(repeal),
    year     = as_factor(year),
    fip      = as_factor(fip),
    fa       = as_factor(fa),
    younger2 = case_when(age == 20 ~ 1, TRUE ~ 0),
    yr2      = as_factor(case_when(repeal == 1 & younger2 == 1 ~ 1, TRUE ~ 0)),
    wm       = as_factor(case_when(wht == 1 & male == 1 ~ 1, TRUE ~ 0)),
    wf       = as_factor(case_when(wht == 1 & male == 0 ~ 1, TRUE ~ 0)),
    bm       = as_factor(case_when(wht == 0 & male == 1 ~ 1, TRUE ~ 0)),
    bf       = as_factor(case_when(wht == 0 & male == 0 ~ 1, TRUE ~ 0))
  )

regddd <- abortion %>%
  filter(bf == 1 & (age == 20 | age ==25)) %>%
  lm_robust(lnr ~ repeal*year + acc + ir + pi + alcohol + crack + poverty + income
  ↪   + ur,
            data = ., weights = totpop, clusters = fip)
```

Figure 68 shows the DDD estimated coefficients for the treated cohort relative to a slightly older 25- to 29-year-old cohort. It's possible that the 25- to 29-year-old cohort is too close in age to function as a satisfactory within-state control; if those age 20–24 have sex with those who are age 25–29, for instance, then SUTVA is violated. There are other age groups, though, that you can try in place of the 25- to 29-year-olds, and I encourage you to do it for both the experience and the insights you might gleam.

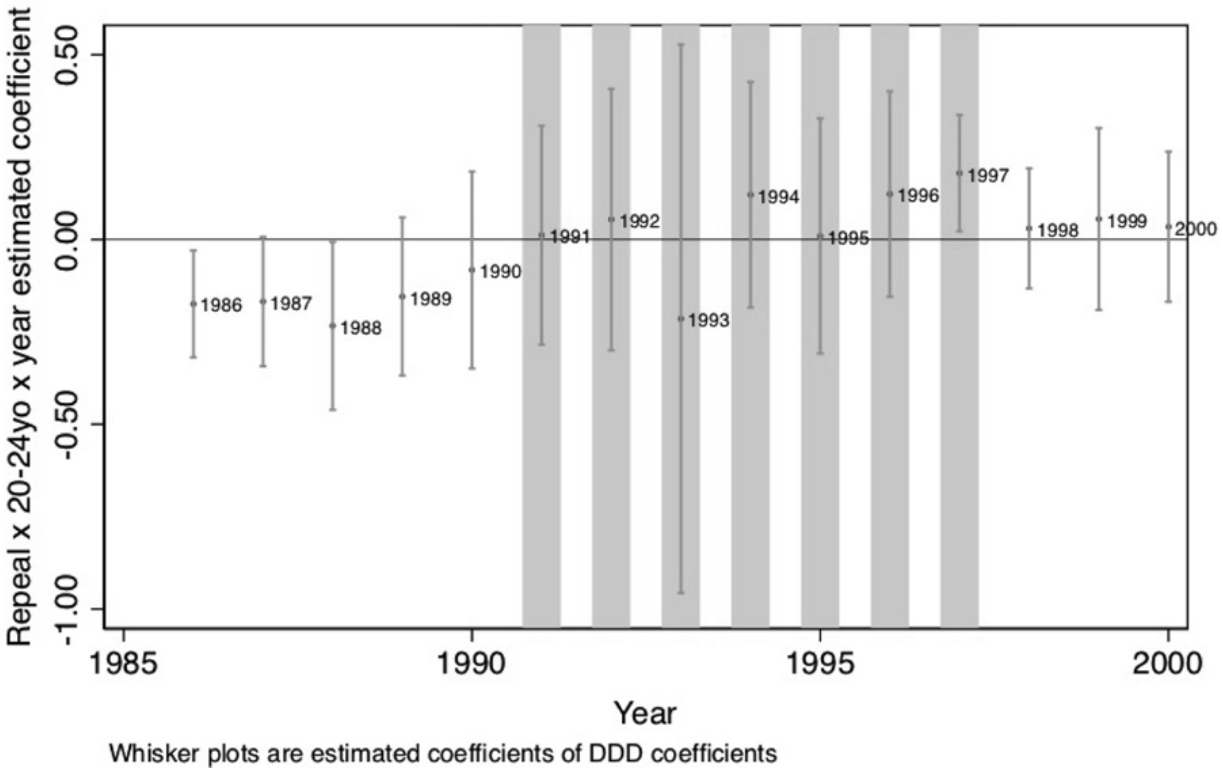


Figure 68. Coefficients and standard errors from DDD regression equation for the 20-to 24-year-olds vs. 25- to 29-year-olds.

But let's back up and remember the big picture. The abortion legalization hypothesis made a series of predictions about where negative parabolic treatment effects should appear in the data. And while we found some initial support, when we exploited more of those predictions, the results fell apart. A fair interpretation of this exercise is that our analysis does *not* support the abortion legalization hypothesis. Figure 68 shows several point estimates at nearly zero, and standard errors so large as to include both positive and negative values for these interactions.

I included this analysis because I wanted to show you the power of a theory with numerous unusual yet testable predictions. Imagine for a moment if a parabola had showed up for all age groups in precisely the years predicted by the theory. Wouldn'twe *have* to update our priors about the abortion legalization selection hypothesis? With predictions so narrow, what else could be causing it? It's precisely because the predictions are so specific, though, that we are able to reject the abortion legalization hypothesis, at least for

gonorrhea. *Placebos as critique*. Since the fundamental problem of causal inference blocks our direct observation of causal effects, we rely on many direct and indirect pieces of evidence to establish credible causality. And as I said in the previous section on DDD, one of those indirect pieces of evidence is placebo analysis. The reasoning goes that if we find, using our preferred research design, effects where there shouldn't be, then maybe our original findings weren't credible in the first place. Using placebo analysis within your own work has become an essential part of empirical work for this reason.

But another use of placebo analysis is to evaluate the credibility of popular estimation strategies themselves. This kind of use helps improve a literature by uncovering flaws in a research design which can then help stimulate the creation of stronger methods and models. Let's take two exemplary studies that accomplished this well: Auld and Grootendorst [2004] and Cohen-Cole and Fletcher [2008].

To say that the Becker and Murphy [1988] "rational addiction" model has been influential would be an understatement. It has over 4,000 cites and has become one of the most common frameworks in health economics. It created a cottage industry of empirical studies that persists to this day. Alcohol, tobacco, gambling, even sports, have all been found to be "rationally addictive" commodities and activities using various empirical approaches.

But some researchers cautioned the research community about these empirical studies. Rogeberg [2004] critiqued the theory on its own grounds, but I'd like to focus on the empirical studies based on the theory. Rather than talk about any specific paper, I'd like to provide a quote from Melberg [2008], who surveyed researchers who had written on rational addiction:

> A majority of [our] respondents believe the literature is a success story that demonstrates the power of economic reasoning. At the same time, they also believe the empirical evidence is weak, and they disagree both on the type of evidence that would validate the theory and the policy implications. Taken together, this points to an interesting gap. On the one hand, most of the respondents claim that the theory has valuable real world implications. On

the other hand, they do not believe the theory has received empirical support. [1]

Rational addiction should be held to the same empirical standards as in theory. The strength of the model has always been based on the economic reasoning, which economists obviously find compelling. But were the empirical designs flawed? How could we know?

Auld and Grootendorst [2004] is not a test of the rational addiction model. On the contrary, it is an "anti-test" of the empirical rational addiction models common at the time. Their goal was not to evaluate the theoretical rational addiction model, in other words, but rather the empirical rational addiction models themselves. How do they do this? Auld and Grootendorst [2004] used the empirical rational addiction model to evaluate commodities that could not plausibly be considered addictive, such as eggs, milk, orange, and apples. They found that the empirical rational addiction model implied milk was extremely addictive, perhaps one of the most addictive commodities studied.[9] Is it credible to believe that eggs and milk are "rationally addictive" or is it more likely the research designs used to evaluate the rational addiction model were flawed? Auld and Grootendorst [2004] study cast doubt on the empirical rational addiction model, not the theory.

Another problematic literature was the peer-effects literature. Estimating peer effects is notoriously hard. Manski [1993] said that the deep endogeneity of social interactions made the identification of peer effects difficult and possibly even impossible. He called this problem the "mirroring" problem. If "birds of a feather flock together," then identifying peer effects in observational settings may just be impossible due to the profound endogeneities at play.

Several studies found significant network effects on outcomes like obesity, smoking, alcohol use, and happiness. This led many researchers to conclude that these kinds of risk behaviors were "contagious" through peer effects [Christakis and Fowler, 2007]. But these studies did not exploit randomized social groups. The peer groups were purely endogenous. Cohen-Cole and Fletcher [2008]

showed using similar models and data that even attributes that *couldn't* be transmitted between peers—acne, height, and headaches—appeared "contagious" in observational data using the Christakis and Fowler [2007] model for estimation. Note, Cohen-Cole and Fletcher [2008] does not reject the idea of theoretical contagions. Rather, they point out that the Manski critique should guide peer effect analysis if social interactions are endogenous. They provide evidence for this indirectly using placebo analysis.[10]

Compositional change within repeated cross-sections. DD can be applied to repeated cross-sections, as well as panel data. But one of the risks of working with the repeated cross-sections 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 author's study exploited the emergence and immense popularity 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 the emergence of Napster. At first glance, they found that as Internet diffusion increased from 1996 to 2001, spending on music for Internet users fell faster than that for non-Internet users. This was initially evidence that Napster was responsible for the decline, until this was investigated more carefully.

But when we look at Table 76, we see evidence of compositional changes. While music expenditure fell over the treatment period, the demographics of the two groups also changed over this period. For instance, the age of Internet users 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 like an omitted variable bias built into the sample itself caused by time-variant unobservables. Diffusion of the Internet appears to be related to changing samples as younger music fans are early adopters. Identification of causal effects would

need for the treatment itself to be exogenous to such changes in the composition.

Final thoughts. There are a few other caveats I'd like to make before moving on. First, it is important to remember the concepts we learned in the early DAG chapter. In choosing covariates in a DD design, you must resist the temptation to simply load the regression up with a kitchen sink of regressors. You should resist if only because in so doing, you may inadvertently include a collider, and if a collider is conditioned on, it introduces strange patterns that may mislead you and your audience. There is unfortunately no way forward except, again, deep institutional familiarity with both the factors that determined treatment assignment on the ground, as well as economic theory itself. Second, another issue I skipped over entirely is the question of how the outcome is modeled. Very little thought if any is given to how exactly we should model some outcome. Just to take one example, should we use the log or the levels themselves? Should we use the quartic root? Should we use rates? These, it turns, out are critically important because for many of them, the parallel trends assumption needed for identification will not be achieved—even though it will be achieved under some other unknown transformation. It is for this reason that you can think of many DD designs as having a parametric element because you must make strong commitments about the functional form itself. I cannot provide guidance to you on this, except that maybe using the pre-treatment leads as a way of finding parallelism could be a useful guide.

Twoway Fixed Effects with Differential Timing

I have a bumper sticker on my car that says "I love Federalism (for the natural experiments)" (Figure 69). I made these bumper stickers for my students to be funny, and to illustrate that the United States is a never-ending laboratory. Because of state federalism, each US state has been given considerable discretion to govern itself with policies and reforms. Yet, because it is a union of states, US

researchers have access to many data sets that have been harmonized across states, making it even more useful for causal inference.

Table 76. Changes between Internet and non-Internet users over time.

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	0.56	0.79	0.60	0.80	0.64	0.81
Entertainment	0.08	0.32	0.09	0.35	0.14	0.39
Demographics						
Age	40.2	49.0	42.3	49.0	44.1	49.4
Income	$52,887	$30,459	$51,995	$26,189	$49,970	$26,649
High school graduate	0.18	0.31	0.17	0.32	0.21	0.32
Some college	0.37	0.28	0.35	0.27	0.34	0.27
College grad	0.43	0.21	0.45	0.21	0.42	0.20
Manager	0.16	0.08	0.16	0.08	0.14	0.08

Note: Sample means from the Consumer Expenditure Survey.

Figure 69. A bumper sticker for nerds.

Goodman-Bacon [2019] calls the staggered assignment of treatments across geographic units over time the “differential timing” of treatment. What he means is unlike the simple 2 × 2 that we discussed earlier (e.g., New Jersey and Pennsylvania), where treatment units were all treated at the same time, the more common situation is one where geographic units receive treatments at different points in time. And this happens in the United States because each area (state, municipality) will adopt a policy when it wants to, for its own reasons. As a result, the adoption of some treatment will tend to be differentially timed across units.

This introduction of differential timing means there are basically two types of DD designs. There is the 2×2 DD we’ve been

discussing wherein a single unit or a group of units all receive some treatment at the same point in time, like Snow's cholera study or Card and Krueger [1994]. And then there is the DD with differential timing in which groups receive treatment at different points in time, like Cheng and Hoekstra [2013]. We have a very good understanding of the 2×2 design, how it works, why it works, when it works, and when it does not work. But we did not until Goodman-Bacon [2019] have as good an understanding of the DD design with differential timing. So let's get down to business and discuss that now by reminding ourselves of the 2×2 DD that we introduced earlier.

$$\widehat{\delta}_{kU}^{2\times2} = \left(\overline{y}_k^{\text{post}(k)} - \overline{y}_k^{\text{pre}(k)}\right) - \left(\overline{y}_U^{\text{post}(k)} - \overline{y}_U^{\text{pre}(k)}\right)$$

where k is the treatment group, U is the never-treated group, and everything else is self-explanatory. Since this involves sample means, we can calculate the differences manually. Or we can estimate it with the following regression:

$$y_{it} = \beta D_i + \tau \text{ Post}_t + \delta(D_i \times \text{Post}_t) + X_{it} + \varepsilon_{it}$$

But a more common situation you'll encounter will be a DD design with differential timing. And while the decomposition is a bit complicated, the regression equation itself is straightforward:

$$y_{it} = \alpha_0 + \delta D_{it} + X_{it} + \alpha_i + \alpha_t + \epsilon_{it}$$

When researchers estimate this regression these days, they usually use the linear fixed-effectsmodel that I discussed in the previous panel chapter. These linear panel models have gotten the nickname "twoway fixed effects" because they include both time fixed effects and unit fixed effects. Since this is such a popular estimator, it's important we understand exactly what it is doing and what is it not.

Bacon Decomposition theorem. Goodman-Bacon [2019] provides a helpful decomposition of the twoway fixed effects estimate of $\widehat{\delta}$.

Given this is the go-to model for implementing differential timing designs, I have found his decomposition useful. But as there are some other decompositions of twoway fixed effects estimators, such as another important paper by de Chaisemartin and D'Haultfoeuille [2019], I'll call it the Bacon decomposition for the sake of branding.

The punchline of the Bacon decomposition theorem is that the twoway fixed effects estimator is a weighted average of all potential 2×2 DD estimates where weights are both based on group sizes and variance in treatment. Under the assumption of variance weighted common trends (VWCT) and time invariant treatment effects, the variance weighted ATT is a weighted average of all possible ATTs. And under more restrictive assumptions, that estimate perfectly matches the ATT. But that is not true when there are time-varying treatment effects, as time-varying treatment effects in a differential timing design estimated with twoway fixed effects can generate a bias. As such, twoway fixed-effects models may be severely biased, which is echoed in de Chaisemartin and D'Haultfoeuille [2019].

To make this concrete, let's start with a simple example. Assume in this design that there are three groups: an early treatment group *(k)*, a group treated later *(l)*, and a group that is never treated *(U)*. Groups *k* and *l* are similar in that they are both treated but they differ in that *k* is treated earlier than *l*.

Let's say there are 5 periods, and *k* is treated in period 2. Then it spends 40% of its time under treatment, or 0.4. But let's say *l* is treated in period 4. Then it spends 80% of its time treated, or 0.8. I represent this time spent in treatment for a group as D_k = 0.4 and D_l = 0.8. This is important, because the length of time a group spends in treatment determines its treatment variance, which in turn affects the weight that 2×2 plays in the final adding up of the DD parameter itself. And rather than write out 2×2 DD estimator every time, we will just represent each 2×2 as $\widehat{\delta}_{ab}^{2\times2}$ where *a* and *b* are the treatment groups, and *j* is the index notation for any treatment group. Thus if we wanted to know the 2×2 for group *k* compared to group *U*, we would write $\widehat{\delta}_{kU}^{2\times2,}$ or, maybe to save space, just $\widehat{\delta}_{kU}^{k}$.

So, let's get started. First, in a single differential timing design, how many 2 × 2s are there anyway? Turns out there are a lot. To see, let's make a toy example. Let's say there are three timing groups (*a*, *b*, and *c*) and one untreated group *(U)*. Then there are 9 2×2 DDs. They are:

a to b	b to a	c to a
a to c	b to c	c to b
a to U	b to U	c to U

See how it works? Okay, then let's return to our simpler example where there are two timing groups *k* and *l* and one never-treated group. Groups *k* and *l* will get treated at time periods t_k^* and t_l^*. The earlier period before anyone is treated will be called the "pre" period, the period between *k* and *l* treated is called the "mid" period, and the period after *l* is treated is called the "post" period. This will be *much* easier to understand with some simple graphs. Let's look at Figure 70. Recall the definition of a 2×2 DD is

$$\widehat{\delta}_{kU}^{2\times2} = \left(\overline{y}_k^{\text{post}(k)} - \overline{y}_k^{\text{pre}(k)}\right) - \left(\overline{y}_U^{\text{post}(k)} - \overline{y}_U^{\text{pre}(k)}\right)$$

where *k* and *U* are just place-holders for any of the groups used in a 2×2.

Substituting the information in each of the four panels of Figure 70 into the equation will enable you to calculate what each specific 2 ×2 is. But we can really just summarize these into three really important 2 × 2s, which are:

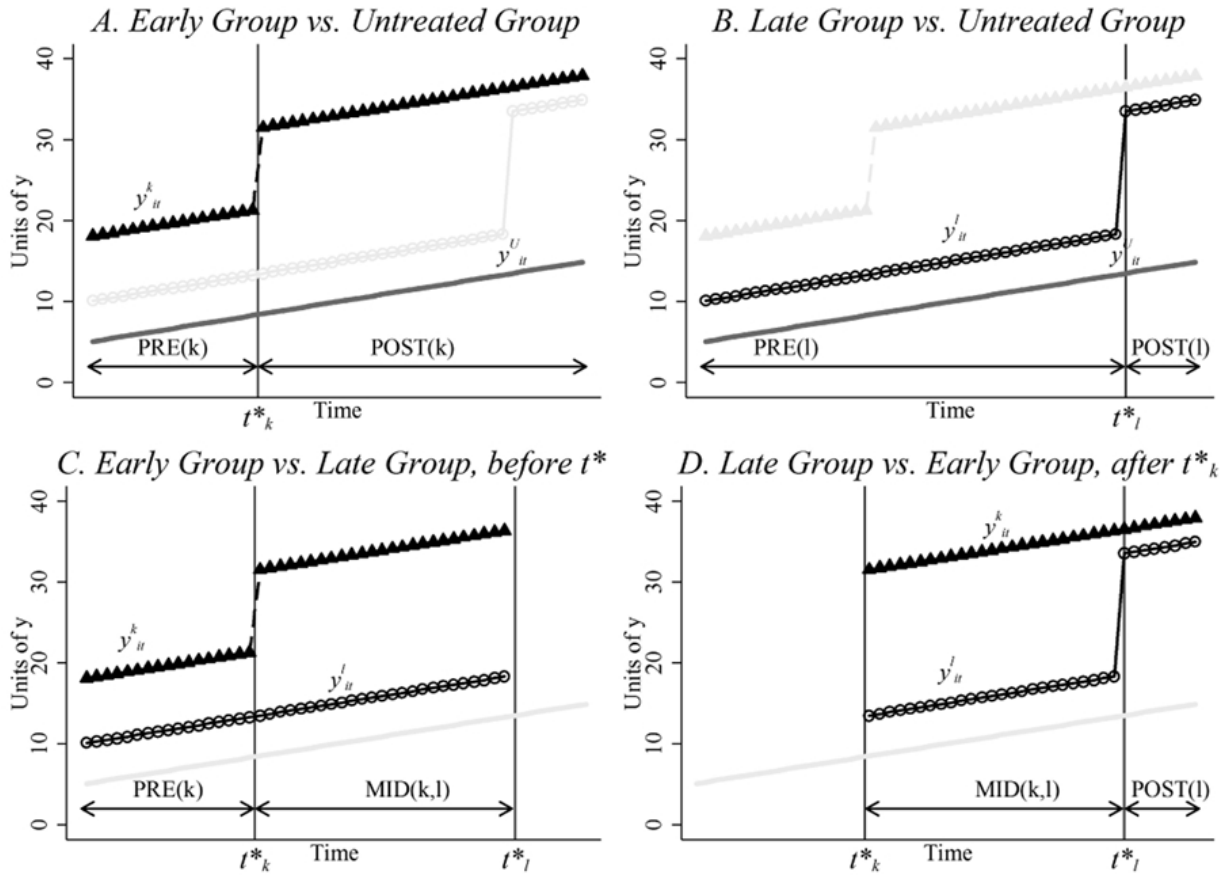


Figure 70. Four 2 × 2 DDs [Goodman-Bacon, 2019]. Reprinted with permission from authors.

$$\widehat{\delta}_{kU}^{2\times2} = \left(\overline{y}_k^{\text{post}(k)} - \overline{y}_k^{\text{pre}(k)}\right) - \left(\overline{y}_U^{\text{post}(k)} - \overline{y}_U^{\text{pre}(k)}\right)$$

$$\widehat{\delta}_{kl}^{2\times2} = \left(\overline{y}_k^{mid(k,l)} - \overline{y}_k^{\text{pre}(k)}\right) - \left(\overline{y}_l^{mid(k,l)} - \overline{y}_l^{\text{pre}(k)}\right)$$

$$\widehat{\delta}_{lk}^{2\times2} = \left(\overline{y}_l^{\text{post}(l)} - \overline{y}_l^{mid(k,l)}\right) - \left(\overline{y}_k^{\text{post}(l)} - \overline{y}_k^{mid(k,l)}\right)$$

where the first 2 × 2 is any timing group compared to the untreated group (*k* or *l*), the second is a group compared to the yet-to-be-treated timing group, and the last is the eventually-treated group compared to the already-treated controls.

With this notation in mind, the DD parameter estimate can be decomposed as follows: $\widehat{\delta}^{DD}$

$$\widehat{\delta}^{DD} = \sum_{k \neq U} s_{kU} \widehat{\delta}^{2\times2}_{kU} + \sum_{k \neq U} \sum_{l > k} s_{kl} \left[\mu_{kl} \widehat{\delta}^{2\times2,k}_{kl} + (1 - \mu_{kl}) \widehat{\delta}^{2\times2,l}_{kl} \right] \qquad (9.1)$$

where the first 2×2 is the *k* compared to *U* and the *l* compared to *U* (combined to make the equation shorter).[11] So what are these weights exactly?

$$s_{ku} = \frac{n_k n_u \overline{D}_k (1 - \overline{D}_k)}{\widehat{Var}(\tilde{D}_{it})}$$

$$s_{kl} = \frac{n_k n_l (\overline{D}_k - \overline{D}_l)(1 - (\overline{D}_k - \overline{D}_l))}{\widehat{Var}(\tilde{D}_{it})}$$

$$\mu_{kl} = \frac{1 - \overline{D}_k}{1 - (\overline{D}_k - \overline{D}_l)}$$

where *n* refers to sample sizes, $\overline{D}_k(1-\overline{D}_k)$ $(\overline{D}_k-\overline{D}_l)(1-(\overline{D}_k-\overline{D}_l))$ expressions refer to variance of treatment, and the final equation is the same for two timing groups.[12]

Two things immediately pop out of these weights that I'd like to bring to your attention. First, notice how "group" variation matters, as opposed to unit-level variation. The Bacon decomposition shows that it's group variation that twoway fixed effects is using to calculate that parameter you're seeking. The more states that adopted a law at the same time, the bigger they influence that final aggregate estimate itself.

The other thing that matters in these weights is *within-group* treatment variance. To appreciate the subtlety of what's implied, ask yourself—how long does a group have to be treated in order to maximize its treatment variance? Define $X = D(1 - D) = D - D^2$, take the derivative of *V* with respect to *D*, set $\frac{dV}{d\overline{D}}$ equal to zero, and solve for $D*$. Treatment variance is maximized when $D = 0.5$. Let's look at three values of *D* to illustrate this.

$$\overline{D} = 0.1; 0.1 \times 0.9 = 0.09$$

$$\overline{D} = 0.4; 0.4 \times 0.6 = 0.24$$

$$\overline{D} = 0.5; 0.5 \times 0.5 = 0.25$$

So what are we learning from this, exactly? Well, what we are learning is that being treated in the *middle* of the panel actually directly influences the numerical value you get when twoway fixed effects are used to estimate the ATT. That therefore means lengthening or shortening the panel can actually change the point estimate purely by changing group treatment variance and nothing more. Isn't that kind of strange though? What criteria would we even use to determine the best length?

But what about the "treated on treated weights," or the s_{kl} weight. That doesn't have a $\bar{D}(1-\bar{D})$ expression. Rather, it has a $(\bar{D}_k - \bar{D}_l)(1-(\bar{D}_k-\bar{D}_l)$ expression. So the "middle" isn't super clear. That's because it isn't the middle of treatment for a single group, but rather it's the middle of the panel for the *difference* in treatment variance. For instance, let's say *k* spends 67% of time treated and *l* spends 15% of time treated. Then $\bar{D}_k - \bar{D}_l = 0.52$ and therefore $0.52 \times 0.48 = 0.2496$, which as we showed is very nearly the max value of the variance as is possible (e.g., 0.25). Think about this for a moment—twoway fixed effects with differential timing weights the 2×2s comparing the two ultimate treatment groups more if the gap in treatment time is close to 0.5.

Expressing the decomposition in potential outcomes. Up to now, we just showed what was inside the DD parameter estimate when using twoway fixed effects: it was nothing more than an "adding up" of all possible 2 × 2s weighted by group shares and treatment variance. But that only tells us what DD is numerically; it does not tell us whether the parameter estimate maps onto a meaningful average treatment effect. To do that, we need to take those sample averages and then use the switching equations replace them with potential

outcomes. This is key to moving from numbers to estimates of causal effects.

Bacon's decomposition theorem expresses the DD coefficient in terms of sample average, making it straightforward to substitute with potential outcomes using a modified switching equation. With a little creative manipulation, this will be revelatory. First, let's define any year-specific ATT as

$$ATT_k(\tau) = E[Y_{it}^1 - Y_{it}^0 \mid k, t = \tau]$$

Next, let's define it over a time window W (e.g., a post-treatment window)

$$ATT_k(\tau) = E[Y_{it}^1 - Y_{it}^0 \mid k, \tau \in W]$$

Finally, let's define differences in average potential outcomes over time as:

$$\Delta Y_k^h(W_1, W_0) = E[Y_{it}^h \mid k, W_1] - E[Y_{it}^h \mid k, W_0]$$

for $h = 0$ (i.e., Y^0) or $h = 1$ (i.e., Y^1)

With trends, differences in mean potential outcomes is non-zero. You can see that in Figure 71.

We'll return to this, but I just wanted to point it out to you so that it would be concrete in your mind when we return to it later.

We can move now from the 2 × 2s that we decomposed earlier directly into the ATT, which is ultimately the main thing we want to know. We covered this earlier in the chapter, but review it again here to maintain progress on my argument. I will first write down the 2×2 expression, use the switching equation to introduce potential outcome notation, and through a little manipulation, find some ATT expression.

$$
\begin{aligned}
\widehat{\delta}_{kU}^{2\times 2} &= \Big(E[Y_j \mid \text{Post}] - E[Y_j \mid \text{Pre}]\Big) - \Big(E[Y_u \mid \text{Post}] - E[Y_u \mid \text{Pre}]\Big) \\
&= \underbrace{\Big(E[Y_j^1 \mid \text{Post}] - E[Y_j^0] \mid \text{Pre}]\Big) - \Big(E[Y_u^0 \mid \text{Post}] - E[Y_u^0 \mid \text{Pre}]\Big)}_{\text{Switching equation}} \\
&+ \underbrace{E[Y_j^0 \mid \text{Post}] - E[Y_j^0 \mid \text{Post}]}_{\text{Adding zero}} \\
&= \underbrace{E[Y_j^1 \mid \text{Post}] - E[Y_j^0 \mid \text{Post}]}_{\text{ATT}} \\
&+ \underbrace{\Big[E[Y_j^0 \mid \text{Post}] - E[Y_j^0 \mid \text{Pre}]\Big] - \Big[E[Y_U^0 \mid \text{Post}] - E[Y_U^0 \mid \text{Pre}]\Big]}_{\text{Non-parallel trends bias in } 2\times 2 \text{ case}}
\end{aligned}
$$

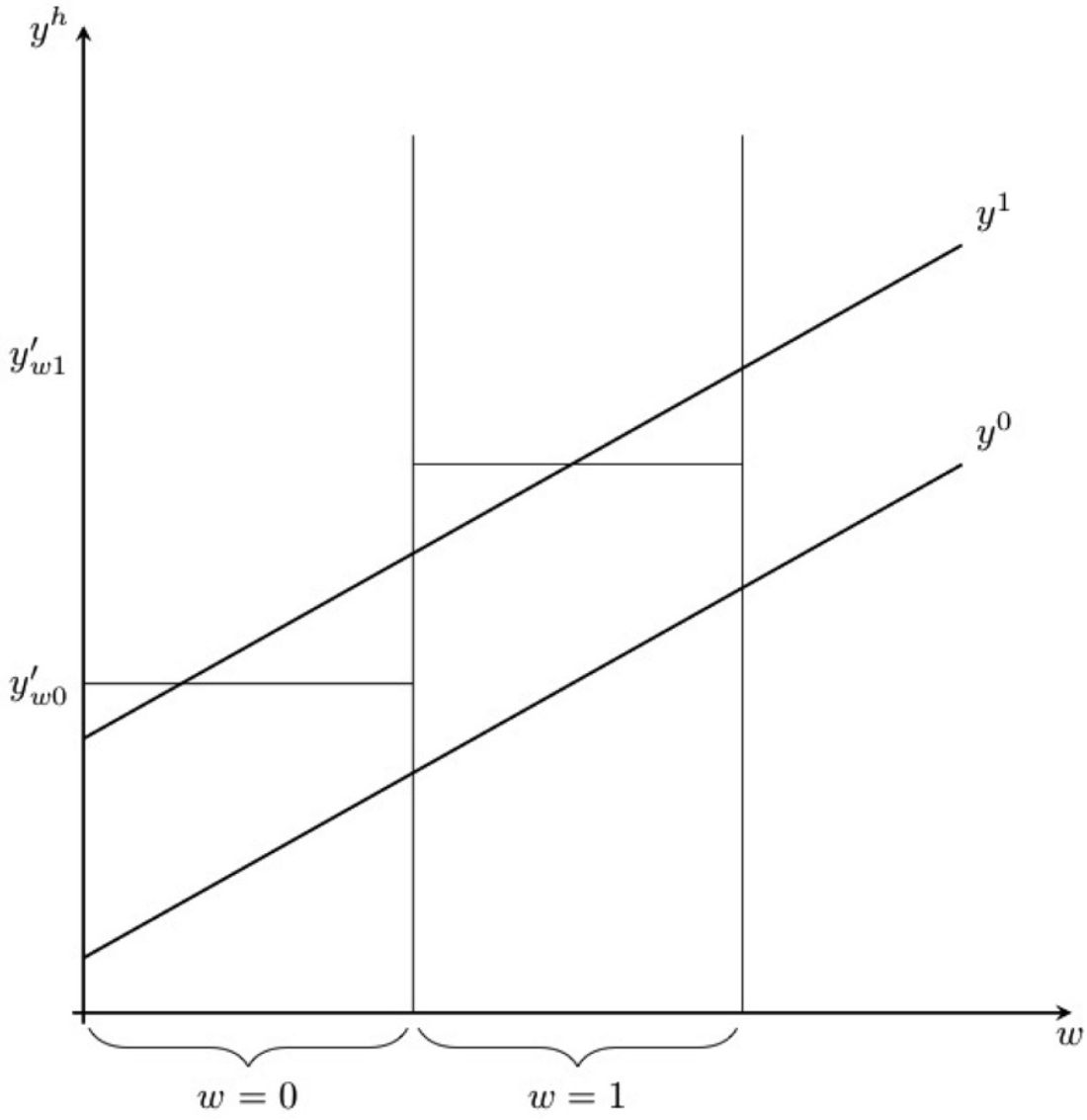


Figure 71. Changing average potential outcomes.

This can be rewritten even more compactly as:

$$\widehat{\delta}_{kU}^{2\times2} = ATT_{\text{Post},j} + \underbrace{\Delta Y^0_{\text{Post,Pre},j} - \Delta Y^0_{\text{Post,Pre},U}}_{\text{Selection bias!}}$$

The 2 × 2 DD can be expressed as the sum of the ATT itself *plus* a parallel trends assumption, and without parallel trends, the estimator is biased. Ask yourself—which of these two differences in the parallel trends assumption is counterfactual,

$\Delta Y^0_{\text{Post,Pre}j}$ or $\Delta Y^0_{\text{Post,Pre}U}$? Which one is observed, in other words, and which one is not observed? Look and see if you can figure it out from this drawing in Figure 72.

Only if these are parallel—the counterfactual trend *and* the observable trend—does the selection bias term zero out and ATT is identified.

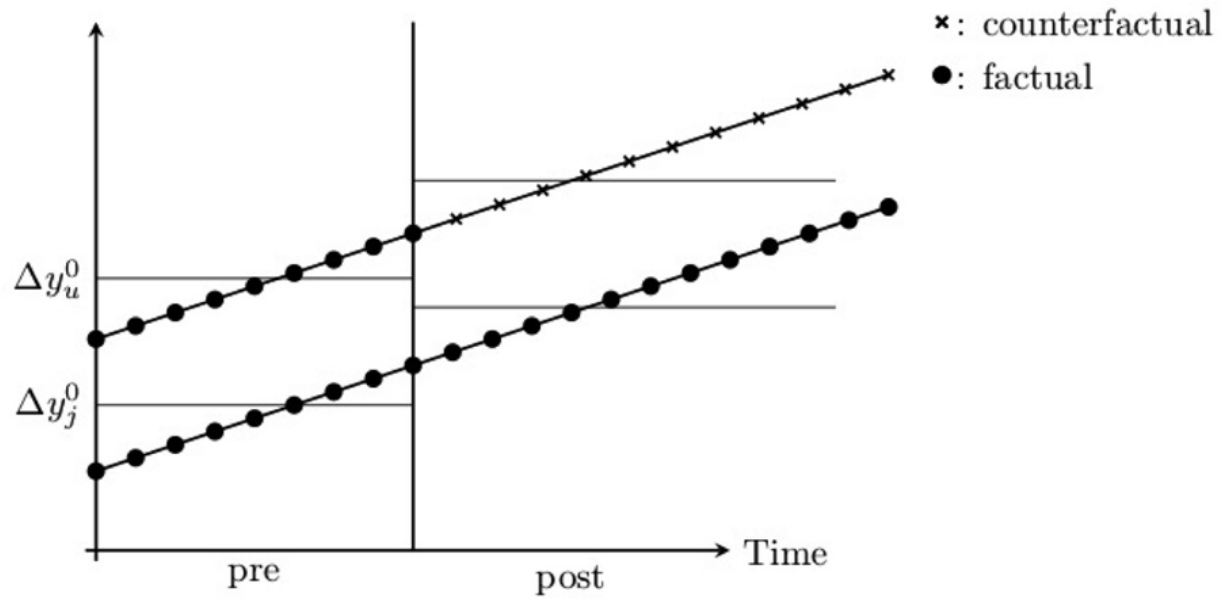


Figure 72. Visualization of parallel trends.

But let's keep looking within the decomposition, as we aren't done. The other two 2×2s need to be defined since they appear in Bacon's decomposition also. And they are:

$$\widehat{\delta}^{2\times2}_{kU} = ATT_k\text{Post} + \Delta Y^0_k(\text{Post}(k),\text{Pre}(k)) - \Delta Y^0_U(\text{Post}(k),\text{Pre}) \quad (9.2)$$

$$\widehat{\delta}^{2\times2}_{kl} = ATT_k(MID) + \Delta Y^0_k(MID,\text{Pre}) - \Delta Y^0_l(MID,\text{Pre}) \quad (9.3)$$

These look the same because you're always comparing the treated unit with an untreated unit (though in the second case it's just that they haven't been treated *yet*).

But what about the 2 × 2 that compared the late groups to the already-treated earlier groups? With a lot of substitutions like we did we get:

$$\widehat{\delta}_{lk}^{2\times2} = ATT_{l,\text{Post}(l)} + \underbrace{\Delta Y_l^0(\text{Post}(l), MID) - \Delta Y_k^0(\text{Post}(l), MID)}_{\text{Parallel-trends bias}} - \underbrace{(ATT_k(\text{Post}) - ATT_k(Mid))}_{\text{Heterogeneity in time bias!}} \tag{9.4}$$

I find it interesting our earlier decomposition of the simple difference in means into *ATE* + selection bias + heterogeneity treatment effects bias resembles the decomposition of the late to early 2×2 DD.

The first line is the *ATT* that we desperately hope to identify. The selection bias zeroes out insofar as Y^0 for *k* and *l* has the same parallel trends from *mid* to *post* period. And the treatment effects bias in the third line zeroes out *so long as* there are constant treatment effects for a group *over time*. But if there is heterogeneity in time for a group, then the two *ATT* terms will not be the same, and therefore will not zero out.

But we can sign the bias if we are willing to assume monotonicity, which means the *mid* term is smaller in absolute value than the *post* term. Under monotonicity, the interior of the parentheses in the third line is positive, and therefore the bias is negative. For positive ATT, this will bias the effects towards zero, and for negative ATT, it will cause the estimated ATT to become even more negative.

Let's pause and collect these terms. The decomposition formula for DD is:

$$\widehat{\delta}^{DD} = \sum_{k\neq U} s_{kU}\widehat{\delta}_{kU}^{2\times2} + \sum_{k\neq U}\sum_{l>k} s_{kl}\left[\mu_{kl}\widehat{\delta}_{kl}^{2\times2,k} + (1-\mu_{kl})\widehat{\delta}_{kl}^{2\times2,l}\right]$$

We will substitute the following three expressions into that formula.

$$\widehat{\delta}_{kU}^{2\times2} = ATT_k(\text{Post}) + \Delta Y_l^0(\text{Post},\text{Pre}) - \Delta Y_U^0(\text{Post},\text{Pre})$$

$$\widehat{\delta}_{kl}^{2\times2,k} = ATT_k(\text{Mid}) + \Delta Y_l^0(\text{Mid},\text{Pre}) - \Delta Y_l^0(\text{Mid},\text{Pre})$$

$$\widehat{\delta}_{lk}^{2\times2,l} = ATT_l\,\text{Post}(l) + \Delta Y_l^0(\text{Post}(l),\text{MID}) - \Delta Y_k^0(\text{Post}(l),\text{MID}) - (ATT_k(\text{Post}) - ATT_k(\text{Mid}))$$

Substituting all three terms into the decomposition formula is a bit overwhelming, so let's simplify the notation. The estimated DD parameter is equal to:

$$p\lim \widehat{\delta}_{n\rightarrow\infty}^{DD} = VWATT + VWCT - \Delta ATT \tag{9.5}$$

In the next few sections, I discuss each individual element of this expression.

Variance weighted ATT. We begin by discussing the variance weighted average treatment effect on the treatment group, or *VWATT*. Its unpacked expression is:

$$VWATT = \sum_{k\neq U} \sigma_{kU} ATT_k(\text{Post}(k)) \tag{9.6}$$

$$+ \sum_{k\neq U}\sum_{l>k} \sigma_{kl}\left[\mu_{kl} ATT_k(\text{MID}) + (1-\mu_{kl}) ATT_l(POST(l))\right] \tag{9.7}$$

where σ is like s, only population terms not samples. Notice that the VWATT simply contains the three ATTs identified above, each of which was weighted by the weights contained in the decomposition formula. While these weights sum to one, that weighting is irrelevant if the ATT are identical.[13]

When I learned that the DD coefficient was a weighted average of all individual 2 × 2s, I was not terribly surprised. I may not have intuitively known that the weights were based on group shares and treatment variance, but I figured it was probably a weighted average nonetheless. I did not have that same experience, though, when I

worked through the other two terms. I now turn to the other two terms: the VWCT and the *ATT*.

Variance weighted common trends. VWCT stands for variance weighted common trends. This is just the collection of non-parallel-trends biases we previously wrote out, but notice—identification requires *variance weighted* common trends to hold, which is actually a bit weaker than we thought before with identical trends. You get this with identical trends, but what Goodman-Bacon [2019] shows us is that *technically* you don't need identical trends because the weights can make it hold even if we don't have exact parallel trends. Unfortunately, this is a bit of a pain to write out, but since it's important, I will.

$$
\begin{aligned}
VWCT \quad &= \sum_{k \neq U} \sigma_{kU} \Big[\Delta Y_k^0(\text{Post}(k), \text{Pre}) - \Delta Y_U^0(\text{Post}(k), \text{Pre}) \Big] \\
&+ \sum_{k \neq U} \sum_{l > k} \sigma_{kl} \Big[\mu_{kl} \{ \Delta Y_k^0(Mid, \text{Pre}(k)) - \Delta Y_l^0(\text{Mid}, \text{Pre}(k)) \} \\
&+ (1 - \mu_{kl}) \{ \Delta Y_l^0(\text{Post}(l), \text{Mid}) - \Delta Y_k^0(\text{Post}(l), \text{Mid}) \} \Big] \quad (9.8)
\end{aligned}
$$

Notice that the VWCT term simply collects all the non-parallel-trend biases from the three 2×2s. One of the novelties, though, is that the non-parallel-trend biases are also weighted by the same weights used in the VATT.

This is actually a new insight. On the one hand, there are a lot of terms we need to be zero. On the other hand, it's ironically a *weaker* identifying assumption strictly identical commontrends as the weights can technically correct for unequal trends. VWCT will zero out with exact parallel trends and in those situations where the weights adjust the trends to zero out. This is good news (sort of).

ATT heterogeneity within time bias. When we decomposed the simple difference in mean outcomes into the sum of the ATE, selection bias, and heterogeneous treatment effects bias, it really

wasn't a huge headache. That was because if the ATT differed from the ATU, then the simple difference in mean outcomes became the sum of ATT and selection bias, which was still an interesting parameter. But in the Bacon decomposition, ATT heterogeneity over time introduces bias that is not so benign. Let's look at what happens when there is time-variant within-group treatment effects.

$$\Delta ATT = \sum_{k \neq U} \sum_{l > k} (1 - \mu_{kl}) \Big[ATT_k(\text{Post}(l) - ATT_k(\text{Mid})) \Big] \qquad (9.9)$$

Heterogeneity in the ATT has two interpretations: you can have heterogeneous treatment effects *across* groups, and you can have heterogeneous treatment effects *within* groups over time. The *ATT* is concerned with the latter only. The first case would be heterogeneity *across units* but not within groups. When there is heterogeneity across groups, then the VWATT is simply the average over group-specific ATTs weighted by a function of sample shares and treatment variance. There is no bias from this kind of heterogeneity.[14]

But it's the second case—when *ATT* is constant across units but heterogeneous within groups over time—that things get a little worrisome. Time-varying treatment effects, even if they are identical across units, generate cross-group heterogeneity because of the differing post-treatment windows, and the fact that earlier-treated groups are serving as controls for later-treated groups. Let's consider a case where the counterfactual outcomes are identical, but the treatment effect is a linear break in the trend (Figure 73). For instance, $Y_{it}^1 = Y_{it}^0 + \theta(t - t_1^* + 1)$ similar to Meer and West [2016].

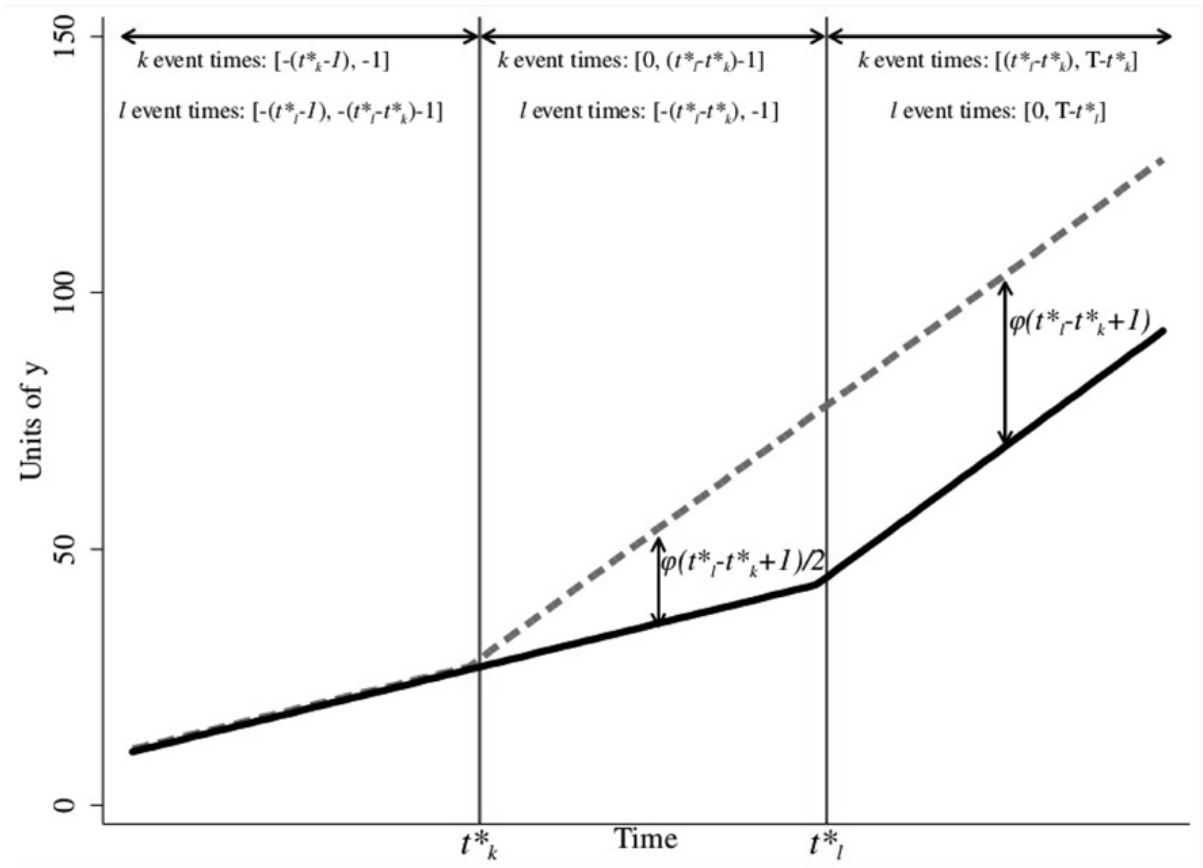


Figure 73. Within-group heterogeneity in the *ATT*. Goodman-Bacon, A. (2019). "Difference-in-Differences with Variation in Treatment Timing." Unpublished Manuscript. Permission from author.

Notice how the first 2×2 uses the later group as its control in the middle period, *but* in the late period, the later-treated group is using the earlier treated as its control. When is this a problem?

It's a problem if there are a lot of those 2× 2s or if their weights are large. If they are negligible portions of the estimate, then even if it exists, then given their weights are small (as group shares are also an important piece of the weighting not just the variance in treatment) the bias may be small. But let's say that doesn't hold. Then what is going on? The effect is biased because the control group is experiencing a trend in outcomes (e.g., heterogeneous treatment effects), and this bias feeds through to the later 2×2 according to the size of the weights, $(1-\mu_{kl})$. We will need to correct for this if our plan is to stick with the twoway fixed effects estimator.

Now it's time to use what we've learned. Let's look at an interesting and important paper by Cheng and Hoekstra [2013] to both learn more about a DD paper and replicate it using event studies and the Bacon decomposition.

Castle-doctrine statutes and homicides. Cheng and Hoekstra [2013] evaluated the impact that a gun reform had on violence and to illustrate various principles and practices regarding differential timing. I'd like to discuss those principles in the context of this paper. This next section will discuss, extend, and replicate various parts of this study.

Trayvon Benjamin Martin was a 17-year-old African-American young man when George Zimmerman shot and killed him in Sanford, Florida, on February 26, 2012. Martin was walking home alone from a convenience store when Zimmerman spotted him, followed him from a distance, and reported him to the police. He said he found Martin's behavior "suspicious," and though police officers urged Zimmerman to stay back, Zimmerman stalked and eventually provoked Martin. An altercation occurred and Zimmerman fatally shot Martin. Zimmerman claimed self-defense and was nonetheless charged with Martin's death. A jury acquitted him of second-degree murder and of manslaughter.

Zimmerman's actions were interpreted by the jury to be legal because in 2005, Florida reformed when and where lethal self-defense could be used. Whereas once lethal self-defense was only legal inside the home, a new law, "Stand Your Ground," had extended that right to other public places. Between 2000 and 2010, twenty-one states explicitly expanded the castle-doctrine statute by extending the places outside the home where lethal force could be legally used.[15] These states had removed a long-standing tradition in the common law that placed the duty to retreat from danger on the victim. After these reforms, though, victims no longer had a duty to retreat in public places if they felt threatened; they could retaliate in lethal self-defense.

Other changes were also made. In some states, individuals who used lethal force outside the home were *assumed* to be reasonably

afraid. Thus, a prosecutor would have to prove fear was not reasonable, allegedly an almost impossible task. Civil liability for those acting under these expansions was also removed. As civil liability is a lower threshold of guilt than criminal guilt, this effectively removed the remaining constraint that might keep someone from using lethal force outside the home.

From an economic perspective, these reforms lowered the cost of killing someone. One could use lethal self-defense in situations from which they had previously been barred. And as there was no civil liability, the expected cost of killing someone was now lower. Thus, insofar as people are sensitive to incentives, then depending on the elasticities of lethal self-defense with respect to cost, we expect an increase in lethal violence for the marginal victim. The reforms may have, in other words, caused homicides to rise.

One can divide lethal force into true and false positives. The true positive use of lethal force would be those situations in which, had the person not used lethal force, he or she would have been murdered. Thus, the true positive case of lethal force is simply a transfer of one life (the offender) for another (the defender). This is tragic, but official statistics would not record an net increase in homicides relative to the counterfactual—only which person had been killed. But a false positive causes a net increase in homicides relative to the counterfactual. Some arguments can escalate unnecessarily, and yet under common law, the duty to retreat would have defused the situation before it spilled over into lethal force. Now, though, under these castle-doctrine reforms, that safety valve is removed, and thus a killing occurs that would not have in counterfactual, leading to a net increase in homicides.

But that is not the only possible impact of the reforms—deterrence of violence is also a possibility under these reforms. In Lott and Mustard [1997], the authors found that concealed-carry laws reduced violence. They suggested this was caused by deterrence—thinking someone may be carrying a concealed weapon, the rational criminal is deterred from committing a crime. Deterrence dates back to Becker [1968] and Jeremy Bentham before him. Expanding the arenas where lethal force could be used could also deter crime.

Since this theoretical possibility depends crucially on key elasticities, which may in fact be zero, deterrence from expanding where guns can be used to kill someone is ultimately an empirical question.

Cheng and Hoekstra [2013] chose a difference-in-differences design for their project where the castle doctrine law was the treatment and timing was differential across states. Their estimating equation was

$$Y_{it} = \alpha + \delta D_{it} + \gamma X_{it} + \sigma_i + \tau_t + \varepsilon_{it}$$

where D_{it} is the treatment parameter. They estimated this equation using a standard twoway fixed effects model as well as count models. Ordinarily, the treatment parameter will be a 0 or 1, but in Cheng and Hoekstra [2013], it's a variable ranging from 0 to 1, because some states get the law change mid-year. So if they got the law in July, then D_{it} equals 0 before the year of adoption, 0.5 in the year of adoption and 1 thereafter. The X_{it} variable included a particular kind of control that they called "region-by-year fixed effects," which was a vector of dummies for the census region to which the state belonged interacted with each year fixed effect. This was done so that explicit counterfactuals were forced to come from within the same census region.[16] As the results are not dramatically different between their twoway fixed effects and count models, I will tend to emphasize results from the twoway fixed effects.

The data they used is somewhat standard in crime studies. They used the FBI Uniform Crime Reports Summary Part I files from 2000 to 2010. The FBI Uniform Crime Reports is a harmonized data set on eight "index" crimes collected from voluntarily participating police agencies across the country. Participation is high and the data goes back many decades, making it attractive for many contemporary questions regarding the crime policy. Crimes were converted into rates, or "offenses per 100,000 population."

Cheng and Hoekstra [2013] rhetorically open their study with a series of simple placebos to check whether the reforms were spuriously correlated with crime trends more generally. Since

oftentimes many crimes are correlated because of unobserved factors, this has some appeal, as it rules out the possibility that the laws were simply being adopted in areas where crime rates were already rising. For their falsifications they chose motor vehicle thefts and larcenies, neither of which, they reasoned, should be credibly connected to lowering the cost of using lethal force in public.

There are so many regression coefficients in Table 77 because applied microeconomists like to report results under increasingly restrictive models. In this case, each column is a new regression with additional controls such as additional fixed-effects specifications, time-varying controls, a one-year lead to check on the pre-treatment differences in outcomes, and state-specific trends. As you can see, many of these coefficients are very small, and because they are small, even large standard errors yield a range of estimates that are still not very large.

Next they look at what they consider to be crimes that might be deterred if policy created a credible threat of lethal retaliation in public: burglary, robbery, and aggravated assault.

Insofar as castle doctrine has a deterrence effect, then we would expect a *negative* effect of the law on offenses. But all of the regressions shown in Table 78 are actually positive, and very few are significant even still. So the authors conclude they cannot detect any deterrence—which does not mean it didn't happen; just that they cannot reject the null for large effects.

Now they move to their main results, which is interesting because it's much more common for authors to lead with their main results. But the rhetoric of this paper is somewhat original in that respect. By this point, the reader has seen a lot of null effects from the laws and may be wondering, "What's going on? This law isn't spurious and isn't causing deterrence. Why am I reading this paper?"

The first thing the authors did was show a series of figures showing the *raw data* on homicides for treatment and control states. This is always a challenge when working with differential timing, though. For instance, approximately twenty states adopted a castle-doctrine law from 2005 to 2010, but *not at the same time*. So how are you going to show this visually? What is the pre-treatment

period, for instance, for the *control group* when there is differential timing? If one state adopts

Table 77. Falsification Tests: The effect of castle doctrine laws on larceny and motor vehicle theft.

	OLS – Weighted by State Population					
	1	2	3	4	5	6
Panel A. Larceny	**Log(Larceny Rate)**					
Castle Doctrine Law	0.00300	−0.00600	−0.00910	−0.0858	−0.00401	−0.00284
	(0.0161)	(0.0147)	(0.0139)	(0.0139)	(0.0128)	(0.0180)
0 to 2 years before adoption of castle doctrine law				0.00112		
				(0.0105)		
Observation	550	550	550	550	550	550
Panel B. Motor Vehicle Theft	**Log(Motor Vehicle Theft Rate)**					
Castle Doctrine Law	0.0517	−0.0389	−0.0252	−0.0294	−0.0165	−0.00708
	(0.0563)	(0.448)	(0.0396)	(0.0469)	(0.0354)	(0.0372)
0 to 2 years before adoption of castle doctrine law					−0.00896	
					(0.0216)	
Observation	550	550	550	550	550	550
State and year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Region-by-year fixed effects		Yes	Yes	Yes	Yes	Yes
Time-varying controls			Yes	Yes	Yes	Yes
Controls for larceny or motor theft					Yes	
State-specific linear time trends						Yes

Notes: Each column in each panel represents a separate regression. The unit of observation is state-year. Robust standard errors are clustered at the state level. Time-varying controls include policing and incarceration rates, welfare and public assistance spending, median income, poverty rate, unemployment rate, and demographics. *Significant at the 10 percent level. **Significant at the 5 percent level. ***Significant at the 1 percent level.

Table 78. The deterrence effects of castle-doctrine laws: Burglary, robbery, and aggravated assault.

	OLS – Weighted by State Population					
	1	2	3	4	5	6
Panel A. Burglary	**Log(Burglary Rate)**					
Castle-doctrine law	0.0780***	0.0290	0.0223	0.0181	0.0327*	0.0237
	(0.0255)	(0.0236)	(0.0223)	(0.0265)	(0.0165)	(0.0207)
0 to 2 years before adoption of castle-doctrine law				−0.009606		
				(0.0133)		
Panel B. Robbery	**Log(Robbery Rate)**					
Castle-doctrine law	0.0408	0.0344	0.0262	0.0197	0.0376**	0.0515*
	(0.0254)	(0.0224)	(0.0229)	(0.0257)	(0.0181)	(0.0274)
0 to 2 years before adoption of castle-doctrine law					−0.0138	
					(0.0153)	
Panel C. Aggravated Assault	**Log(Aggravated Assault Rate)**					
Castle-doctrine law	0.0434	0.0397	0.0372	0.0330	0.0424	0.0414
	(0.0387)	(0.0407)	(0.0319)	(0.0367)	(0.0291)	(0.0285)
0 to 2 years before adoption of castle-doctrine law					−0.00897	
					(0.0147)	
Observation	550	550	550	550	550	550
State and year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Region-by-year fixed effects		Yes	Yes	Yes	Yes	Yes
Time-varying controls		Yes	Yes	Yes		Yes
Controls for larceny or motor theft				Yes		
State-specific linear time trends						Yes

(*Continued*)

Notes: Each column in each panel represents a separate regression. The unit of observation is state-year. Robust standard errors are clustered at the state level. Time-varying controls include policing and incarceration rates, welfare and public assistance spending, median income, poverty rate, unemployment rate, and demographics. *Significant at the 10 percent level. **Significant at the 5 percent level. ***Significant at the 1 percent level.

in 2005, but another in 2006, then what precisely is the pre- and post-treatment for the control group? So that's a bit of a challenge, and yet if you stick with our guiding principle that causal inference studies desperately need data visualization of the main effects, your job is to solve it with creativity and honesty to make beautiful figures. Cheng and Hoekstra [2013] could've presented regression coefficients on leads and lags, as that is very commonly done, but knowing these authors firsthand, their preference is to give the reader pictures of the raw data to be as transparent as possible. Therefore, they showed multiple figures where each figure was a "treatment group" compared to all the "never-treated" units. Figure 74 shows the Florida case.

Notice that before the passage of the law, the offenses are fairly flat for treatment and control. Obviously, as I've emphasized, this is not a direct *test* of the parallel-trends assumption. Parallel trends in the pre-treatment period are neither necessary nor sufficient. The identifying assumption, recall, is that of variance-weighted common trends, which are entirely based on parallel counterfactual trends, not pretreatment trends. But researchers use parallel pre-treatment trends like a hunch that the counterfactual trends would have been parallel. In one sense, parallel pre-treatment rules out some obvious spurious factors that we should be worried about, such as the law adoption happening around the timing of a change, even if that's simply nothing more than seemingly spurious factors like rising homicides. But that's clearly not happening here—homicides weren't diverging from controls pre-treatment. They were following a similar trajectory before Florida passed its law and *only then* did the trends converge. Notice that after 2005, which is when the law occurs, there's a sizable jump in homicides. There are additional figures like

this, but they all have this set up—they show a treatment *group* over time compared to the same "never-treated" group.

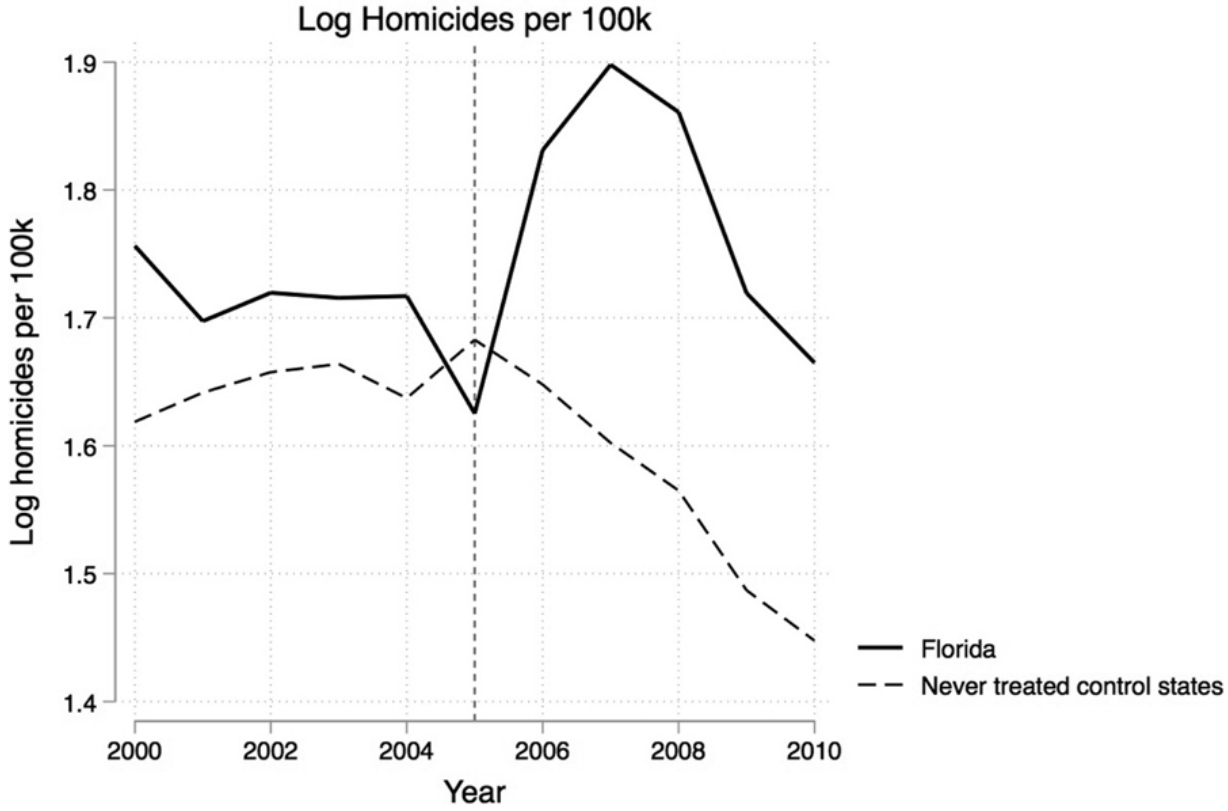


Figure 74. Raw data of log homicides per 100,000 for Florida versus never-treated control states.

Insofar as the cost of committing lethal force has fallen, then we expect to see more of it, which implies a positive coefficient on the estimated δ term assuming the heterogeneity bias we discussed earlier doesn't cause the twoway fixed effects estimated coefficient to flip signs. It should be different from zero both statistically and in a meaningful magnitude. They present four separate types of specifications—three using OLS, one using negative binomial. But I will only report the weighted OLS regressions for the sake of space.

There's a lot of information in Table 79, so let's be sure not to get lost. First, all coefficients are positive and similar in magnitude—between 8% and 10% increases in homicides. Second, three of the four panels are almost entirely significant. It appears that the bulk of

their evidence suggests the castle-doctrine statute caused an increase in homicides around 8%.

Table 79. The effect of castle-doctrine laws on homicide.

Panel A. Homicide OLS–Weights	Log(Homicide rate) 1	2	3	4	5	6
Castle-doctrine law	0.0801**	0.0946***	0.0937***	0.0955**	0.0985***	0.100**
	(0.0342)	(0.0279)	(0.0290)	(0.0367)	(0.0299)	(0.0388)
0 to 2 years before adoption of castle-doctrine law					0.00398	
					(0.0222)	
Observation	550	550	550	550	550	550
State and year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Region-by-year fixed effects		Yes	Yes	Yes	Yes	Yes
Time-varying controls			Yes	Yes	Yes	Yes
Controls for larceny or motor theft					Yes	
State-specific linear time trends						Yes

Notes: Each column in each panel represents a separate regression. The unit of observation is state-year. Robust standard errors are clustered at the state level. Time-varying controls include policing and incarceration rates, welfare and public assistance spending, median income, poverty rate, unemployment rate, and demographics. *Significant at the 10 percent level. **Significant at the 5 percent level. ***Significant at the 1 percent level.

Table 80. Randomization inference averages [Cheng and Hoekstra, 2013].

Method	Average estimate	Estimates larger than actual estimate
Weighted OLS	−0.003	0/40
Unweighted OLS	0.001	1/40
Negative binomial	0.001	0/40

Not satisfied, the authors implemented a kind of randomization inference-based test. Specifically, they moved the eleven-year panel back in time covering 1960–2009 and estimated forty placebo “effects” of passing castle doctrine one to forty years earlier. When they did this, they found that the average effect from this exercise was essentially zero. Those results are summarized here. It appears there is something statistically unusual about the actual treatment

profile compared to the placebo profiles, because the actual profile yields effect sizes larger than all but one case in any of the placebo regressions run.

Cheng and Hoekstra [2013] found no evidence that castle-doctrine laws deter violent offenses, but they did find that it increased homicides. An 8% net increase in homicide rates translates to around six hundred additional homicides per year across the twenty-one adopting states. Thinking back to to the killing of Trayvon Martin by George Zimmerman, one is left to wonder whether Trayvon might still be alive had Florida not passed Stand Your Ground. This kind of counterfactual reasoning can drive you crazy, because it is unanswerable—we simply don't know, cannot know, and never will know the answer to counterfactual questions. The fundamental problem of causal inference states that we need to know what would have happened that fateful night without Stand Your Ground and compare that with what happened with Stand Your Ground to know what can and cannot be placed at the feet of that law. What we do know is that under certain assumptions related to the DD design, homicides were on net around 8%–10% higher than they would've been when compared against explicit counterfactuals. And while that doesn't answer every question, it suggests that a nontrivial number of deaths can be blamed on laws similar to Stand Your Ground.

Replicating Cheng and Hoekstra [2013], sort of. Now that we've discussed Cheng and Hoekstra [2013], I want to replicate it, or at least do some work on their data set to illustrate certain things that we've discussed, like event studies and the Bacon decomposition. This analysis will be slightly different from what they did, though, because their policy variable was on the interval [0, 1] rather than being a pure dummy. That's because they carefully defined their policy variable according to the month in which the law was passed (e.g., June) divided by a total of 12 months. So if a state passed the last in June, then they would assign a 0.5 in the first year, and a 1 thereafter. While there's nothing wrong with that approach, I am going to use a dummy because it makes the event studies a bit

easier to visualize, and the Bacon decomposition only works with dummy policy variables.

First, I will replicate his main homicide results from Panel A, column 6, of Figure 74.

STATA

castle_1.do

```
use https://github.com/scunning1975/mixtape/raw/master/castle.dta, clear
set scheme cleanplots
* ssc install bacondecomp

* define global macros
global crime1 jhcitizen_c jhpolice_c murder homicide  robbery assault burglary
    ↪ larceny motor robbery_gun_r
global demo blackm_15_24 whitem_15_24 blackm_25_44 whitem_25_44
    ↪ //demographics
global lintrend trend_1-trend_51 //state linear trend
global region r20001-r20104  //region-quarter fixed effects
global exocrime l_larceny l_motor // exogenous crime rates
global spending l_exp_subsidy l_exp_pubwelfare
global xvar l_police unemployrt poverty l_income l_prisoner l_lagprisoner $demo
    ↪ $spending

label variable post "Year of treatment"
xi: xtreg l_homicide i.year $region $xvar $lintrend post [aweight=popwt], fe
    ↪ vce(cluster sid)

```

R

castle_1.R

```r
library(bacondecomp)
library(tidyverse)
library(haven)
library(lfe)

read_data <- function(df)
{
  full_path <- paste("https://raw.github.com/scunning1975/mixtape/master/",
                     df, sep = "")
  df <- read_dta(full_path)
  return(df)
}

castle <- read_data("castle.dta")

#--- global variables
crime1 <- c("jhcitizen_c", "jhpolice_c",
            "murder", "homicide",
            "robbery", "assault", "burglary",
            "larceny", "motor", "robbery_gun_r")

demo <- c("emo", "blackm_15_24", "whitem_15_24",
          "blackm_25_44", "whitem_25_44")

# variables dropped to prevent colinearity
dropped_vars <- c("r20004", "r20014",
                  "r20024", "r20034",
                  "r20044", "r20054",
                  "r20064", "r20074",
                  "r20084", "r20094",
                  "r20101", "r20102", "r20103",
                  "r20104", "trend_9", "trend_46",
                  "trend_49", "trend_50", "trend_51"
)

lintrend <- castle %>%
    select(starts_with("trend")) %>%
  colnames %>%
  # remove due to colinearity
  subset(.,! . %in% dropped_vars)
```

(continued)

R *(continued)*

```r

region <- castle %>%
  select(starts_with("r20")) %>%
  colnames %>%
  # remove due to colinearity
  subset(.,! . %in% dropped_vars)


exocrime <- c("l_lacerny", "l_motor")
spending <- c("l_exp_subsidy", "l_exp_pubwelfare")


xvar <- c(
  "blackm_15_24", "whitem_15_24", "blackm_25_44", "whitem_25_44",
  "l_exp_subsidy", "l_exp_pubwelfare",
  "l_police", "unemployrt", "poverty",
  "l_income", "l_prisoner", "l_lagprisoner"
)

law <- c("cdl")

dd_formula <- as.formula(
  paste("l_homicide ~ ",
        paste(
          paste(xvar, collapse = " + "),
          paste(region, collapse = " + "),
          paste(lintrend, collapse = " + "),
          paste("post", collapse = " + "), sep = " + "),
        "| year + sid | 0 | sid"
  )
)

#Fixed effect regression using post as treatment variable
dd_reg <- felm(dd_formula, weights = castle$popwt, data = castle)
summary(dd_reg)


```

Here we see the main result that castle doctrine expansions led to an approximately 10% increase in homicides. And if we use the post-dummy, which is essentially equal to 0 unless the state had fully covered castle doctrine expansions, then the effect is more like 7.6%.

But now, I'd like to go beyond their study to implement an event study. First, we need to define pre-treatment leads and lags. To do this, we use a "time_til" variable, which is the number of years until or after the state received the treatment. Using this variable, we then create the leads (which will be the years prior to treatment) and lags (the years post-treatment).

STATA

castle_2.do

```
* Event study regression with the year of treatment (lag0) as the omitted category.
xi: xtreg l_homicide  i.year $region lead9 lead8 lead7 lead6 lead5 lead4 lead3 lead2 lead1 lag1-lag5 [aweight=popwt], fe vce(cluster sid)
```

R

castle_2.R

```
castle <- castle %>%
  mutate(
    time_til = year - treatment_date,
    lead1 = case_when(time_til == -1 ~ 1, TRUE ~ 0),
    lead2 = case_when(time_til == -2 ~ 1, TRUE ~ 0),
    lead3 = case_when(time_til == -3 ~ 1, TRUE ~ 0),
    lead4 = case_when(time_til == -4 ~ 1, TRUE ~ 0),
    lead5 = case_when(time_til == -5 ~ 1, TRUE ~ 0),
    lead6 = case_when(time_til == -6 ~ 1, TRUE ~ 0),
    lead7 = case_when(time_til == -7 ~ 1, TRUE ~ 0),
    lead8 = case_when(time_til == -8 ~ 1, TRUE ~ 0),
    lead9 = case_when(time_til == -9 ~ 1, TRUE ~ 0),

    lag0 = case_when(time_til == 0 ~ 1, TRUE ~ 0),
    lag1 = case_when(time_til == 1 ~ 1, TRUE ~ 0),
    lag2 = case_when(time_til == 2 ~ 1, TRUE ~ 0),
    lag3 = case_when(time_til == 3 ~ 1, TRUE ~ 0),
    lag4 = case_when(time_til == 4 ~ 1, TRUE ~ 0),
    lag5 = case_when(time_til == 5 ~ 1, TRUE ~ 0)
  )
```

(continued)

R *(continued)*

```
event_study_formula <- as.formula(
  paste("l_homicide ~ + ",
        paste(
          paste(region, collapse = " + "),
          paste(paste("lead", 1:9, sep = ""), collapse = " + "),
          paste(paste("lag", 1:5, sep = ""), collapse = " + "), sep = " + "),
        "| year + state | 0 | sid"
  ),
)

event_study_reg <- felm(event_study_formula, weights = castle$popwt, data = castle)
summary(event_study_reg)
```

Our omitted category is the year of treatment, so all coefficients are with respect to that year. You can see from the coefficients on the leads that they are not statistically different from zero prior to treatment, except for leads 8 and 9, which may be because there are only three states with eight years prior to treatment, and one state with nine years prior to treatment. But in the years prior to treatment, leads 1 to 6 are equal to zero and statistically insignificant, although they do technically have large confidence intervals. The lags, on the other hand, are all positive and not too dissimilar from one another except for lag 5, which is around 17%.

Now it is customary to plot these event studies, so let's do that now. I am going to show you an easy way and a longer way to do this. The longer way gives you ultimately more control over what exactly you want the event study to look like, but for a fast and dirty method, the easier way will suffice. For the easier way, you will need to install a program in Stata called coefplot, written by Ben Jann, author of estout.[17]

STATA

castle_3.do

```
* Plot the coefficients using coefplot
* ssc install coefplot

coefplot, keep(lead9 lead8 lead7 lead6 lead5 lead4 lead3 lead2 lead1 lag1 lag2
    ↪ lag3 lag4 lag5) xlabel(, angle(vertical)) yline(0) xline(9.5) vertical
    ↪ msymbol(D) mfcolor(white) ciopts(lwidth(*3) lcolor(*.6)) mlabel
    ↪ format(%9.3f) mlabposition(12) mlabgap(*2) title(Log Murder Rate)
```

R

castle_3.R

```

# order of the coefficients for the plot
plot_order <- c("lead9", "lead8", "lead7",
                "lead6", "lead5", "lead4", "lead3",
                "lead2", "lead1", "lag1",
                "lag2", "lag3", "lag4", "lag5")

# grab the clustered standard errors
# and average coefficient estimates
# from the regression, label them accordingly
# add a zero'th lag for plotting purposes
leadslags_plot <- tibble(
  sd = c(event_study_reg$cse[plot_order], 0),
  mean = c(coef(event_study_reg)[plot_order], 0),
  label = c(-9,-8,-7,-6, -5, -4, -3, -2, -1, 1,2,3,4,5, 0)
)

# This version has a point-range at each
# estimated lead or lag
# comes down to stylistic preference at the
# end of the day!
leadslags_plot %>%
  ggplot(aes(x = label, y = mean,
             ymin = mean-1.96*sd,
             ymax = mean+1.96*sd)) +
  geom_hline(yintercept = 0.035169444, color = "red") +
  geom_pointrange() +
```

(continued)

R *(continued)*

```
  theme_minimal() +
  xlab("Years before and after castle doctrine expansion") +
  ylab("log(Homicide Rate)") +
  geom_hline(yintercept = 0,
             linetype = "dashed") +
  geom_vline(xintercept = 0,
             linetype = "dashed")
```

Let's look now at what this command created. As you can see in Figure 75, eight to nine years prior to treatment, treatment states

have significantly lower levels of homicides, but as there are so few states that even have these values (one with −9 and three with −8), we may want to disregard the relevance of these negative effects if for no other reason than that there are so few units in the dummy and we know from earlier that that can lead to very high overrejection rates [MacKinnon and Webb, 2017]. Instead, notice that for the six years prior to treatment, there is virtually no difference between the treatment states and the control states.

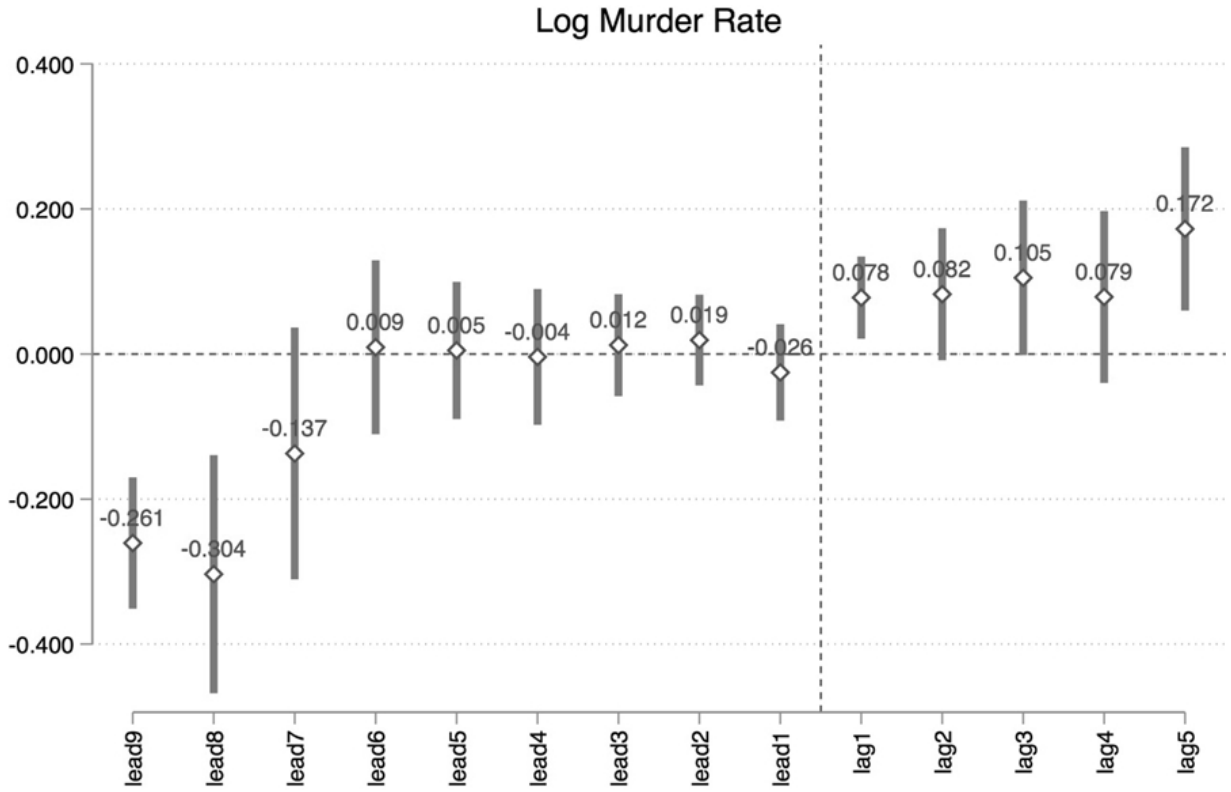


Figure 75. Homicide event-study plots using coefplot. Cheng and Hoekstra [2013].

But, after the year of treatment, that changes. Log murders begin rising, which is consistent with our post dummy that imposed zeros on all pre-treatment leads and required that the average effect post-treatment be a constant.

I promised to show you how to make this graph in a way that gave more flexibility, but you should be warned, this is a bit more cumbersome.

STATA

castle_4.do

```
xi: xtreg l_homicide  i.year $region $xvar $lintrend post [aweight=popwt], fe
    ↪  vce(cluster sid)

local DDL = _b[post]
local DD : display %03.2f _b[post]
local DDSE : display %03.2f _se[post]
local DD1 = -0.10

xi: xtreg l_homicide  i.year $region lead9 lead8 lead7 lead6 lead5 lead4 lead3
    ↪  lead2 lead1 lag1-lag5 [aweight=popwt], fe vce(cluster sid)

outreg2 using "./eventstudy_levels.xls", replace keep(lead9 lead8 lead7 lead6
    ↪  lead5 lead4 lead3 lead2 lead1 lag1-lag5) noparen noaster addstat(DD, `DD',
    ↪  DDSE, `DDSE')


*Pull in the ES Coefs
xmluse "./eventstudy_levels.xls", clear cells(A3:B32) first
replace VARIABLES = subinstr(VARIABLES,"lead","",.)
replace VARIABLES = subinstr(VARIABLES,"lag","",.)
quietly destring _all, replace ignore(",")
replace VARIABLES = -9 in 2
replace VARIABLES = -8 in 4
replace VARIABLES = -7 in 6
replace VARIABLES = -6 in 8
replace VARIABLES = -5 in 10
replace VARIABLES = -4 in 12
replace VARIABLES = -3 in 14
replace VARIABLES = -2 in 16

```

(continued)

STATA *(continued)*

```
27
28
29  replace VARIABLES = -1 in 18
30  replace VARIABLES = 1 in 20
31  replace VARIABLES = 2 in 22
32  replace VARIABLES = 3 in 24
33  replace VARIABLES = 4 in 26
34  replace VARIABLES = 5 in 28
35  drop in 1
36  compress
37  quietly destring _all, replace ignore(",")
38  compress
39
40
41
42  ren VARIABLES exp
43  gen b = exp<.
44  replace exp = -9 in 2
45  replace exp = -8 in 4
46  replace exp = -7 in 6
47  replace exp = -6 in 8
48  replace exp = -5 in 10
49  replace exp = -4 in 12
50  replace exp = -3 in 14
51  replace exp = -2 in 16
52  replace exp = -1 in 18
53  replace exp = 1 in 20
54  replace exp = 2 in 22
55  replace exp = 3 in 24
56  replace exp = 4 in 26
57  replace exp = 5 in 28
58
59  * Expand the dataset by one more observation so as to include the comparison
    ↪  year
60  local obs =_N+1
61  set obs `obs'
62  for var _all: replace X = 0 in `obs'
63  replace b = 1 in `obs'
64  replace exp = 0 in `obs'
65  keep exp l_homicide b
```

(continued)

STATA *(continued)*

```
set obs 30
foreach x of varlist exp l_homicide b {
      replace `x'=0 in 30
      }
reshape wide l_homicide, i(exp) j(b)


* Create the confidence intervals
cap drop *lb* *ub*
gen lb = l_homicide1 - 1.96*l_homicide0
gen ub = l_homicide1 + 1.96*l_homicide0


* Create the picture
set scheme s2color
#delimit ;
twoway (scatter l_homicide1 ub lb exp ,
            lpattern(solid dash dash dot dot solid solid)
            lcolor(gray gray gray red blue)
            lwidth(thick medium medium medium medium thick thick)
            msymbol(i i i i i i i i i i i i i i i) msize(medlarge medlarge)
            mcolor(gray black gray gray red blue)
            c(l l l l l l l l l l l l l l l)
            cmissing(n n n n n n n n n n n n n n n)
            xline(0, lcolor(black) lpattern(solid))
            yline(0, lcolor(black))
            xlabel(-9 -8 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 , labsize(medium))
            ylabel(, nogrid labsize(medium))
            xsize(7.5) ysize(5.5)
            legend(off)
            xtitle("Years before and after castle doctrine expansion",
            ↪  size(medium))
            ytitle("Log Murders ", size(medium))
            graphregion(fcolor(white) color(white) icolor(white) margin(zero))
            yline(`DDL', lcolor(red) lwidth(thick)) text(`DD1' -0.10 "DD
            ↪  Coefficient = `DD' (s.e. = `DDSE')")
            )
            ;

#delimit cr;
```

R

castle_4.R

```
1
2  # This version includes
3  # an interval that traces the confidence intervals
4  # of your coefficients
5  leadslags_plot %>%
6    ggplot(aes(x = label, y = mean,
7               ymin = mean-1.96*sd,
8               ymax = mean+1.96*sd)) +
9    # this creates a red horizontal line
10   geom_hline(yintercept = 0.035169444, color = "red") +
11   geom_line() +
12   geom_point() +
13   geom_ribbon(alpha = 0.2) +
14   theme_minimal() +
15   # Important to have informative axes labels!
16   xlab("Years before and after castle doctrine expansion") +
17   ylab("log(Homicide Rate)") +
18   geom_hline(yintercept = 0) +
19   geom_vline(xintercept = 0)
```

You can see the figure that this creates in Figure 76. The difference between coefplot and this twoway command connects the event-study coefficients with lines, whereas coefplot displayed them as coefficients hanging in the air. Neither is right or wrong; I merely wanted you to see the differences for your own sake and to have code that you might experiment with and adapt to your own needs.

But the thing about this graph is that the leads are imbalanced. There's only one state, for instance, in the ninth lead, and there's only three in the eighth lead. So I'd like you to do two modifications to this. First, I'd like you to replace the sixth lead so that it is now equal to leads 6–9. In other words, we will force these late adopters to have the same coefficient as those with six years until treatment. When you do that, you should get Figure 77.

Next, let's balance the event study by dropping the states who only show up in the seventh, eighth, and ninth leads.[18] When you do this, you should get Figure 78.

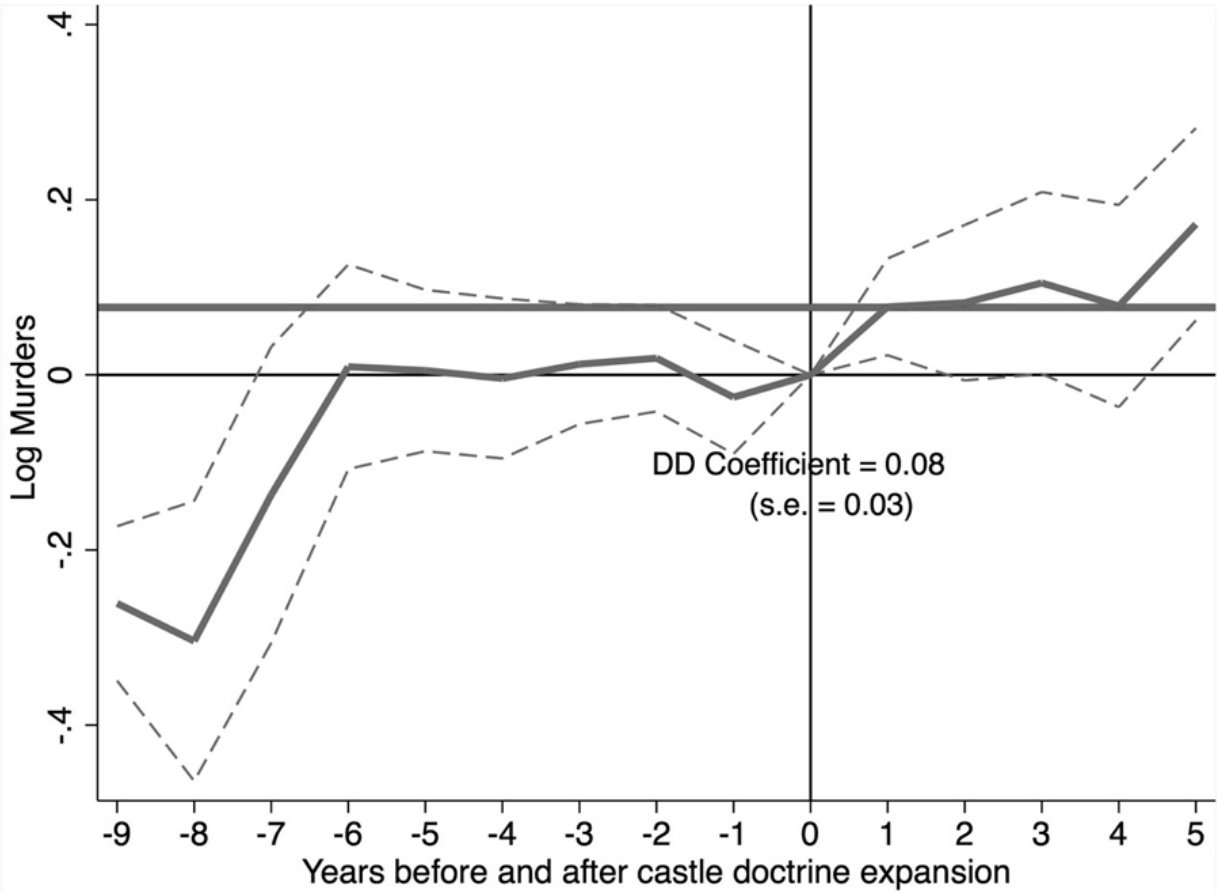


Figure 76. Homicide event study plots created manually with twoway. Cheng and Hoekstra [2013].

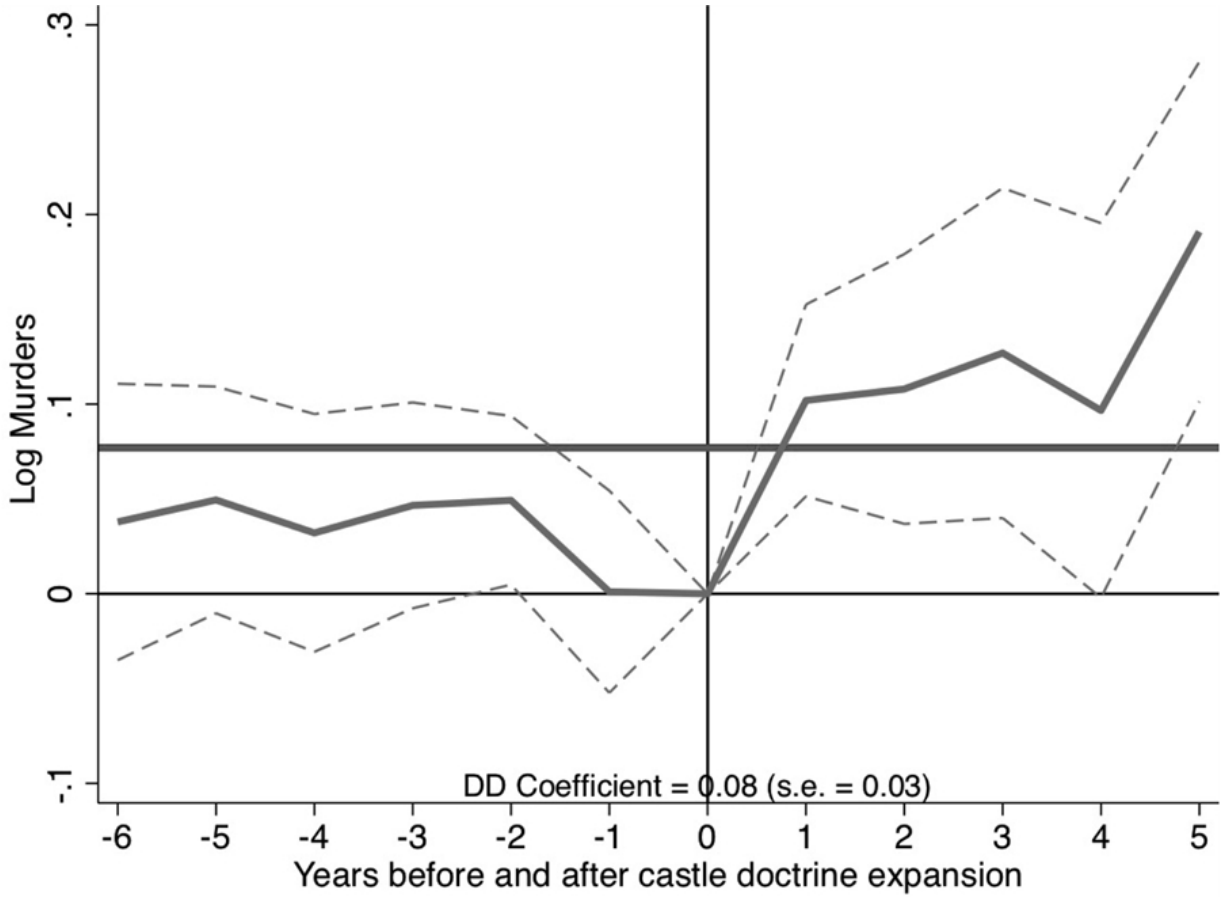


Figure 77. Homicide event-study plots using twoway. Cheng and Hoekstra [2013].

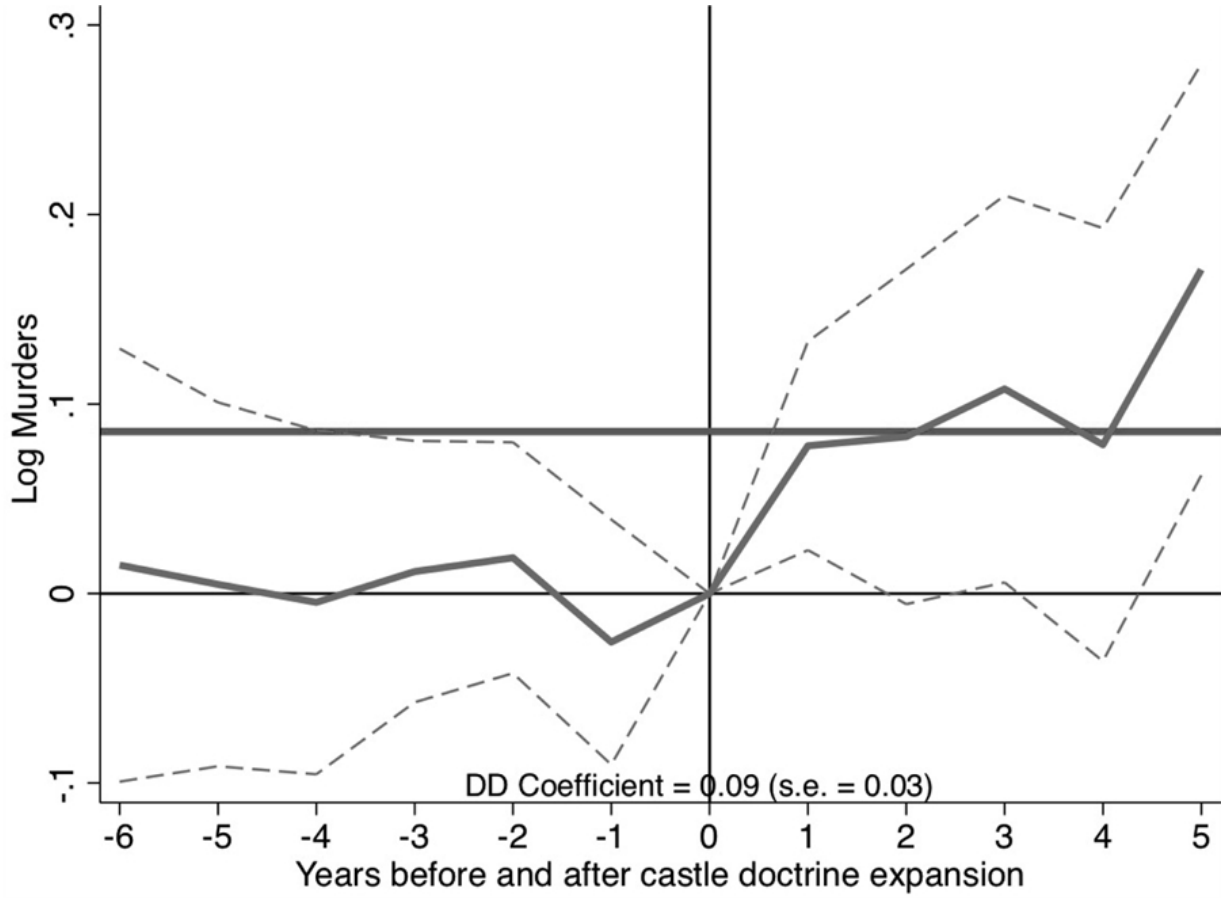


Figure 78. Homicide event-study plots using twoway. Cheng and Hoekstra [2013].

If nothing else, exploring these different specifications and cuts of the data can help you understand just how confident you should be that prior to treatment, treatment and control states genuinely were pretty similar. And if they weren't similar, it behooves the researcher to at minimum provide some insight to others as to why the treatment and control groups were dissimilar in levels. Because after all—if they were different in levels, then it's entirely plausible they would be different in their counterfactual trends too because why else are they different in the first place [Kahn-Lang and Lang, 2019].

Bacon decomposition. Recall that we run into trouble using the twoway fixed-effects model in a DD framework insofar as there are heterogeneous treatment effects over time. But the problem here only occurs with those 2 × 2s that use late-treated units compared to

early-treated units. If there are few such cases, then the issue is much less problematic depending on the magnitudes of the weights and the size of the DD coefficients themselves. What we are now going to do is simply evaluate the frequency with which this issue occurs using the Bacon decomposition. Recall that the Bacon decomposition decomposes the twoway fixed effects estimator of the DD parameter into weighted averages of individual 2 × 2s across the four types of 2 × 2s possible. The Bacon decomposition uses a binary treatment variable, so we will reestimate the effect of castle-doctrine statutes on logged homicide rates by coding a state as "treated" if any portion of the year it had a castle-doctrine amendment. We will work with the special case of no covariates for simplicity, though note that the decomposition works with the inclusion of covariates as well [Goodman-Bacon, 2019]. Stata users will need to download -ddtiming- from Thomas Goldring's website, which I've included in the first line.

First, let's estimate the actual model itself using a post dummy equaling one if the state was covered by a castle-doctrine statute that year. Here we find a smaller effect than many of Cheng and Hoekstra's estimates because we do not include their state-year interaction fixed effects strategy among other things. But this is just for illustrative purposes, so let's move to the Bacon decomposition itself. We can decompose the parameter estimate into the three different types of 2 × 2s, which I've reproduced in Table 81.

STATA

castle_5.do

```
1  use https://github.com/scunning1975/mixtape/raw/master/castle.dta, clear
2  * ssc install bacondecomp
3
4  * define global macros
5  global crime1 jhcitizen_c jhpolice_c murder homicide  robbery assault burglary
   ↪  larceny motor robbery_gun_r
6  global demo blackm_15_24 whitem_15_24 blackm_25_44 whitem_25_44
   ↪  //demographics
7  global lintrend trend_1-trend_51 //state linear trend
8  global region r20001-r20104  //region-quarter fixed effects
9  global exocrime l_larceny l_motor // exogenous crime rates
10 global spending l_exp_subsidy l_exp_pubwelfare
11 global xvar l_police unemployrt poverty l_income l_prisoner l_lagprisoner $demo
   ↪  $spending
12 global law cdl
13
14 * Bacon decomposition
15 net install ddtiming, from(https://tgoldring.com/code/)
16 areg l_homicide post i.year, a(sid) robust
17 ddtiming l_homicide post, i(sid) t(year)
18
```

R

castle_5.R

```r
library(bacondecomp)
library(lfe)

df_bacon <- bacon(l_homicide ~ post,
                  data = castle, id_var = "state",
                  time_var = "year")

# Diff-in-diff estimate is the weighted average of
# individual 2x2 estimates
dd_estimate <- sum(df_bacon$estimate*df_bacon$weight)

# 2x2 Decomposition Plot
bacon_plot <- ggplot(data = df_bacon) +
  geom_point(aes(x = weight, y = estimate,
                 color = type, shape = type), size = 2) +
  xlab("Weight") +
  ylab("2x2 DD Estimate") +
  geom_hline(yintercept = dd_estimate, color = "red") +
  theme_minimal() +
  theme(
    legend.title = element_blank(),
    legend.background = element_rect(
      fill="white", linetype="solid"),
    legend.justification=c(1,1),
    legend.position=c(1,1)
  )

bacon_plot

# create formula
bacon_dd_formula <- as.formula(
  'l_homicide ~ post | year + sid | 0 | sid')

# Simple diff-in-diff regression
bacon_dd_reg <- felm(formula = bacon_dd_formula, data = castle)
summary(bacon_dd_reg)

# Note that the estimate from earlier equals the
# coefficient on post
dd_estimate

```

Table 81. Bacon decomposition example.

DD Comparison	Weight	Avg DD Est
Earlier T vs. Later C	0.077	−0.029
Later T vs. Earlier C	0.024	0.046
T vs. Never treated	0.899	0.078

Dep var	**Log(homicide rate)**
Castle-doctrine law	0.069
	(0.034)

Taking these weights, let's just double check that they do indeed add up to the regression estimate we just obtained using our twoway fixed-effects estimator.[19]

$$di(0.077 * -0.029) + (0.024 * 0.046) + (0.899 * 0.078) = 0.069$$

That is our main estimate, and thus confirms what we've been building on, which is that the DD parameter estimate from a twoway fixed-effects estimator is simply a weighted average over the different types of 2 × 2s in any differential design. Furthermore, we can see in the Bacon decomposition that most of the 0.069 parameter estimate is coming from comparing the treatment states to a group of never-treated states. The average DD estimate for that group is 0.078 with a weight of 0.899. So even though there is a later to early 2×2 in the mix, as there always will be with any differential timing, it is small in terms of influence and ultimately pulls down the estimate.

But let's now visualize this as distributing the weights against the DD estimates, which is a useful exercise. The horizontal line in Figure 79 shows the average DD estimate we obtained from our fixed-effects regression of 0.069. But then what are these other graphics? Let's review.

Each icon in the graphic represents a single 2×2 DD. The horizontal axis shows the weight itself, whereas the vertical axis shows the magnitude of that particular 2×2. Icons further to the right therefore will be more influential in the final average DD than those closer to zero.

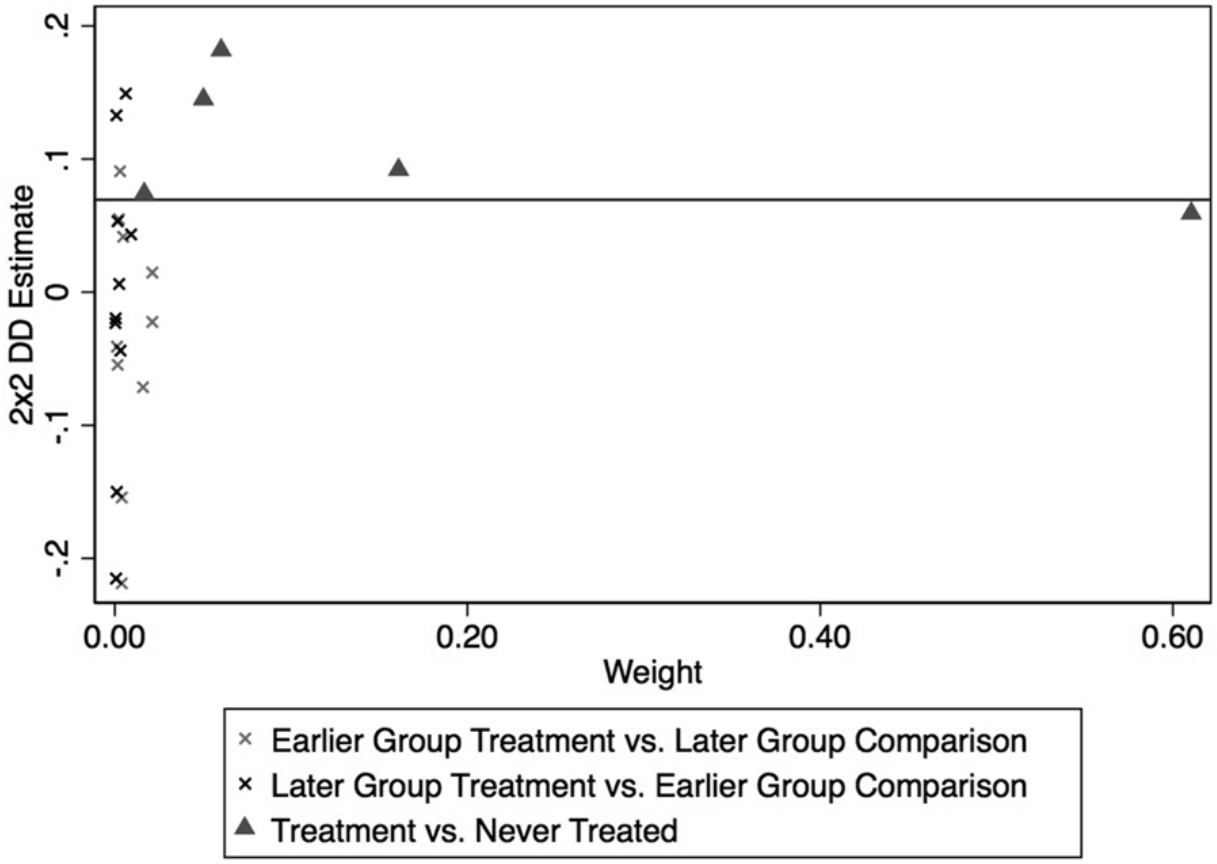


Figure 79. Bacon decomposition of DD into weights and single 2×2s.

There are three kinds of icons here: an early to late group comparison (represented with a light ×), a late to early (dark ×), and a treatment compared to a never-treated (dark triangle). You can see that the dark triangles are all above zero, meaning that each of these 2 × 2s (which correspond to a particular set of states getting the treatment in the same year) is positive. Now they are spread out somewhat—two groups are on the horizontal line, but the rest are higher. What appears to be the case, though, is that the group with the largest weight is really pulling down the parameter estimate and bringing it closer to the 0.069 that we find in the regression.

The future of DD. The Bacon decomposition is an important phase in our understanding of the DD design when implemented using the twoway fixed effects linear model. Prior to this decomposition, we had only a metaphorical understanding of the necessary conditions for identifying causal effects using differential timing with a twoway fixed-effects estimator. We thought that since the 2×2 required parallel trends, that that "sort of" must be what's going on with differential timing too. And we weren't too far off—there is a version of parallel trends in the identifying assumptions of DD using twoway fixed effects with differential timing. But what Goodman-Bacon [2019] also showed is that the weights themselves drove the numerical estimates too, and that while some of it was intuitive (e.g., group shares being influential) others were not (e.g., variance in treatment being influential).

The Bacon decomposition also highlighted some of the unique challenges we face with differential timing. Perhaps no other problem is better highlighted in the diagnostics of the Bacon decomposition than the problematic "late to early" 2×2 for instance. Given any heterogeneity bias, the late to early 2×2 introduces biases *even with variance weighted common trends holding*! So, where to now?

From 2018 to 2020, there has been an explosion of work on the DD design. Much of it is unpublished, and there has yet to appear any real consensus among applied people as to how to handle it. Here I would like to outline what I believe could be a map as you attempt to navigate the future of DD. I have attempted to divide this new work into three categories: weighting, selective choice of "good" 2 × 2s, and matrix completion.

What we know now is that there are two fundamental problems with the DD design. First, there is the issue of weighting itself. The twoway fixed-effects estimator weights the individual 2 × 2s in ways that do not make a ton of theoretical sense. For instance, why do we think that groups at the middle of the panel should be weighted more than those at the end? There's no theoretical reason we should believe that. But as Goodman-Bacon [2019] revealed, that's precisely what twoway fixed effects does. And this is weird because you can change your results simply by adding or subtracting years to

the panel—not just because this changes the 2×2, but also because it changes the variance in treatment itself! So that's weird.[20]

But this is not really the fatal problem, you might say, with twoway fixed-effects estimates of a DD design. The bigger issue was what we saw in the Bacon decomposition—you will inevitably use past treated units as controls for future treated units, or what I called the "late to early 2×2." This happens both in the event study and in the designs modeling the average treatment effect with a dummy variable. Insofar as it takes more than one period for the treatment to be fully incorporated, then insofar as there's substantial weight given to the late to early 2×2s, the existence of heterogeneous treatment effects skews the parameter away from the ATT—maybe even flipping signs![21]

Whereas the weird weighting associated with twoway fixed effects is an issue, it's something you can at least check into because the Bacon decomposition allows you to separate out the 2×2 average DD values from their weights. Thus, if your results are changing by adding years because your underlying 2 × 2s are changing, you simply need to investigate it in the Bacon decomposition. The weights and the 2 × 2s, in other words, are things that can be directly calculated, which can be a source of insight into why twoway fixed effects estimator is finding what it finds.

But the second issue is a different beast altogether. And one way to think of the emerging literature is that many authors are attempting to solve the problem that some of these 2 × 2s (e.g., the late to early 2×2) are problematic. Insofar as they are problematic, can we improve over our static twoway fixed-effects model? Let's take a few of these issues up with examples from the growing literature.

Another solution to the weird weighting twoway fixed-effects problem has been provided by Callaway and Sant'Anna [2019].[22] Callaway and Sant'Anna [2019] approach the DD framework very differently than Goodman-Bacon [2019]. Callaway and Sant'Anna [2019] use an approach that allows them to estimate what they call the group-time average treatment effect, which is just the ATT for a

given group at any point in time. Assuming parallel trends conditional on time-invariant covariates and overlap in a propensity score, which I'll discuss below, you can calculate group ATT by time (relative time like in an event study or absolute time). One unique part of these authors' approach is that it is non-parametric as opposed to regression-based. For instance, under their identifying assumptions, their nonparametric estimator for a group ATT by time is:

$$ATT(g,t) = E\left[\left(\frac{G_g}{E[G_g]} - \frac{\frac{p_g(X)C}{1-p_g(X)}}{E\left[\frac{p_g(X)C}{1-p_g(X)}\right]}\right)(Y_t - Y_{g-1})\right]$$

where the weights, p, are propensity scores, G is a binary variable that is equal to 1 if an individual is first treated in period g, and C is a binary variable equal to one for individuals in the control group. Notice there is no time index, so these C units are the never-treated group. If you're still with me, you should find the weights straightforward. Take observations from the control group as well as group g, and omit the other groups. Then weight up those observations from the control group that have characteristics similar to those frequently found in group g and weight down observations from the control group that have characteristics rarely found in group g. This kind of reweighting procedure guarantees that the covariates of group g and the control group are balanced. You can see principles from earlier chapters making their way into this DD estimation—namely, balance on covariates to create exchangeable units on observables.

But because we are calculating group-specific ATT by time, you end up with a lot of treatment effect parameters. The authors address this by showing how one can take all of these treatment effects and collapse them into more interpretable parameters, such as a larger ATT. All of this is done without running a regression, and therefore avoids some of the unique issues created in doing so.

One simple solution might be to estimate your event-study model and simply take the mean over all lags using a linear combination of

all point estimates [Borusyak and Jaravel, 2018]. Using this method, we in fact find considerably larger effects or nearly twice the size as we get from the simpler static twoway fixed-effects model. This is perhaps an improvement because weights can be large on the long-run effects due to large effects from group shares. So if you want a summary measure, it's better to estimate the event study and then average them after the fact.

Another great example of a paper wrestling with the biases brought up by heterogeneous treatment effects is Sun and Abraham [2020]. This paper is primarily motivated by problems created in event studies, but you can see some of the issues brought up in Goodman-Bacon [2019]. In an event study with differential timing, as we discussed earlier, leads and lags are often used to measure dynamics in the treatment itself. But these can produce causally uninterpretable results because they will assign non-convex weights to cohort-specific treatment effects. Similar to Callaway and Sant'Anna [2019], they propose estimating a group-specific dynamic effect and from those calculate a group specific estimate.

The way I organize these papers in my mind is around the idea of heterogeneity in time, the use of twoway fixed effects, and differential timing. The theoretical insight from all these papers is the coefficients on the static twoway fixed-effects leads and lags will be unintelligible if there is heterogeneity in treatment effects over time. In this sense, we are back in the world that Goodman-Bacon [2019] revealed, in which heterogeneity treatment effect biases create real challenges for the DD design using twoway fixed effects.[23]

Their alternative is estimate a "saturated" model to ensure that the heterogeneous problem never occurs in the first place. The proposed alternative estimation technique is to use an interacted specification that is saturated in relative time indicators as well as cohort indicators. The treatment effect associated with this design is called the interaction-weighted estimator, and using it, the DD parameter is equivalent to the difference between the average change in outcomes for a given cohort in those periods prior to treatment and the average changes for those units that had not been

treated at the time interval. Additionally, this method uses the never-treated units as controls, and thereby avoids the hairy problems noted in Goodman-Bacon [2019] when computing later to early 2 × 2s.[24]

Another paper that attempts to circumvent the weirdness of the regression-based method when there are numerous late to early 2 × 2s is Cengiz et al. [2019]. This is bound to be a classic study in labor for its exhaustive search for detectable repercussions of the minimum wage on low-paying jobs. The authors ultimately find little evidence to support any concern, but how do they come to this conclusion?

Cengiz et al. [2019] take a careful approach by creating separate samples. The authors want to know the impact of minimum-wage changes on low-wage jobs across 138 state-level minimum-wage changes from 1979 to 2016. The authors in an appendix note the problems with aggregating individual DD estimates into a single parameter, and so tackle the problem incrementally by creating 138 separate data sets associated with a minimum-wage event. Each sample has both treatment groups and control groups, but not all units are used as controls. Rather, only units that were not treated within the sample window are allowed to be controls. Insofar as a control is not treated during the sample window associated with a treatment unit, it can be by this criteria used as a control. These 138 estimates are then stacked to calculate average treatment effects. This is an alternative method to the twoway fixed-effects DD estimator because it uses a more stringent criteria for whether a unit can be considered a control. This in turn circumvents the heterogeneity problems that Goodman-Bacon [2019] notes because Cengiz et al. [2019] essentially create 138 DD situations in which controls are always “never-treated” for the duration of time under consideration.

But the last methodology I will discuss that has emerged in the last couple of years is a radical departure from the regression-based methodology altogether. Rather than use a twoway fixed-effects estimator to estimate treatment effects with differential timing, Athey

et al. [2018] propose a machine-learning-based methodology called "matrix completion" for panel data. The estimator is exotic and bears some resemblance to matching imputation and synthetic control. Given the growing popularity of placing machine learning at the service of causal inference, I suspect that once Stata code for matrix completion is introduced, we will see this procedure used more broadly.

Matrix completion for panel data is a machine-learning-based approach to causal inference when one is working explicitly with panel data and differential timing. The application of matrix completion to causal inference has some intuitive appeal given one of the ways that Rubin has framed causality is as a missing data problem. Thus, if we are missing the matrix of counterfactuals, we might explore whether this method from computer science could assist us in recovering it. Imagine we could create two matrices of potential outcomes: a matrix of Y^0 potential outcomes for all panel units over time and Y^1. Once treatment occurs, a unit switches from Y^0 to Y^1 under the switching equation, and therefore the missing data problem occurs. Missingness is simply another way of describing the fundamental problem of causal inference for there will never be a complete set of matrices enabling calculation of interesting treatment parameters given the switching equation only assigns one of them to reality.

Say we are interested in this treatment effect parameter:

$$\widehat{\delta_{ATT}} = \frac{1}{N_T} \sum (Y_{it}^1 - Z_{it}^0)$$

where Y^1 are the observed outcomes in a panel unit at some post-treatment period, Z^0 is the estimated missing elements of the Y^0 matrix for the post-treatment period, and N_T is the number of treatment units. Matrix completion uses the observed elements of the matrix's realized values to predict the missing elements of the Y^0 matrix (missing due to being in the post-treatment period and therefore having switched from Y^0 to Y^1).

Analytically, this imputation is done via something called regularization-based prediction. The objective in this approach is to optimally predict the missing elements by minimizing a convex function of the difference between the observed matrix of Y^0 and the unknown complete matrix Z^0 using nuclear norm regularization. Let denote the row and column indices *(i*, *j)* of the observed entries of the outcomes, then the objective function can be written as

$$\widehat{Z^0} = \underset{Z^0}{\arg\min} \sum_{(i,j)\in\Omega} \frac{(Y_{it}^0 - Z_{it}^0)^2}{|\Omega|} + \Lambda||Z^0||$$

where $||Z^0||$ is the nuclear norm (sum of singular values of $Z0$). The regularization parameter Δ is chosen using tenfold cross validation. Athey et al. [2018] show that this procedure outperforms other methods in terms of root mean squared prediction error.

Unfortunately, at present estimation using matrix completion is not available in Stata. R packages for it do exist, such as the gsynth package, but it has to be adapted for Stata users. And until it is created, I suspect adoption will lag.

Conclusion

America's institutionalized state federalism provides a constantly evolving laboratory for applied researchers seeking to evaluate the causal effects of laws and other interventions. It has 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 "difference-in-differences" brought up 45,000 hits. It is arguably the most common methodology you will use—more than IV or matching or even RDD, despite RDD's greater perceived 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.

But, what we have learned in this chapter is that while there is a current set of identifying assumptions and practices associated with the DD design, differential timing does introduce some thorny challenges that have long been misunderstood. Much of the future of DD appears to be mounting solutions to problems we are coming to understand better, such as the odd weighting of regression itself and problematic 2×2 DDs that bias the aggregate ATT when heterogeneity in the treatment effects over time exists. Nevertheless, DD—and specifically, regression-based DD—is not going away. It is the single most popular design in the applied researcher's toolkit and likely will be for many years to come. Thus it behooves the researcher to study this literature carefully so that they can better protect against various forms of bias.

Notes

1 A simple search on Google Scholar for phrase "difference-in-differences" yields over forty thousand hits.

2 John Snow is one of my personal heroes. He had a stubborn commitment to the truth and was unpersuaded by low-quality causal evidence. That simultaneous skepticism and open-mindedness gave him the willingness to question common sense when common sense failed to provide satisfactory explanations.

3 You'll sometimes see acronyms for difference-in-differences like DD, DiD, Diff-indiff, or even, God forbid, DnD.

4 That literature is too extensive to cite here, but one can find reviews of a great deal of the contemporary literature on minimum wages in Neumark et al. [2014] and Cengiz et al. [2019].

5 James Buchanan won the Nobel Prize for his pioneering work on the theory of public choice. He was not, though, a labor economist, and to my knowledge did not have experience estimating causal effects using explicit counterfactuals with observational data. A Google Scholar search for "James Buchanan minimum wage" returned only one hit, the previously mentioned *Wall Street Journal* letter to the editor. I consider his criticism to be ideologically motivated ad hominem and as such unhelpful in this debate.

6 Financial economics also has a procedure called the event study [Binder, 1998], but the way that event study is often used in contemporary causal inference is nothing more than a difference-in-differences design where,

instead of a single post-treatment dummy, you saturate a model with leads and lags based on the timing of treatment.

7 In the original Cunningham and Cornwell [2013], we estimated models with multiway clustering correction, but the package for this in Stata is no longer supported. Therefore, we will estimate the same models as in Cunningham and Cornwell [2013] using cluster robust standard errors. In all prior analysis, I clustered the standard errors at the state level so as to maintain consistency with this code.

8 There is a third prediction on the 25- to 29-year-olds, but for the sake of space, I only focus on the 20- to 24-year-olds.

9 Milk is ironically my favorite drink, even over IPAs, so I am not persuaded by this anti-test.

10 Breakthroughs in identifying peer effects eventually emerged, but only from studies that serendipitously had randomized peer groups such as Sacerdote [2001], Lyle [2009], Carrell et al. [2019], Kofoed and McGovney [2019], and several others. Many of these papers either used randomized roommates or randomized companies at military academies. Such natural experiments are rare opportunities for studying peer effects for their ability to overcome the mirror problem.

11 All of this decomposition comes from applying the Frisch-Waugh theorem to the underlying twoway fixed effects estimator.

12 A more recent version of Goodman-Bacon [2019] rewrites this weighting but they are numerically the same, and for these purposes, I prefer the weighting scheme discussed in an earlier version of the paper. See Goodman-Bacon [2019] for the equivalence between his two weighting descriptions.

13 Heterogeneity in ATT across k and l is not the source of any biases. Only heterogeneity over time for k or l's ATT introduces bias. We will discuss this in more detail later.

14 Scattering the weights against the individual 2 × 2s can help reveal if the overall coefficient is driven by a few different 2 × 2s with large weights.

15 These laws are called castle-doctrine statutes because the home—where lethal self-defense had been protected—is considered one's castle.

16 This would violate SUTVA insofar as gun violence spills over to a neighboring state when the own state passes a reform.

17 Ben Jann is a valuable contributor to the Stata community for creating several community ado packages, such as -estout- for making tables and -coefplot- for making pictures of regression coefficients.

18 Alex Bartik once recommended this to me.

19 This result is different from Cheng and Hoekstra's because it does not include the region by year fixed effects. I exclude them for simplicity.

20 This is less an issue with event study designs because the variance of treatment indicator is the same for everyone.

21 In all seriousness, it is practically modal in applied papers that utilize a DD design to imagine that dynamic treatment effects are at least plausible ex ante, if not expected. This kind of "dynamic treatment effects" is usually believed as a realistic description of what we think could happen in any policy environment. As such, the biases associated with panel fixed effects model with twoway fixed effects is, to be blunt, scary. Rarely have I seen a study wherein the treatment was merely a one-period shift in size. Even in the Miller et al. [2019] paper, the effect of ACA-led Medicaid expansions was a gradual reduction in annual mortality over time. Figure 60 really is probably a *typical* kind of event study, not an exceptional one.

22 Sant'Anna has been particularly active in this area in producing elegant econometric solutions to some of these DD problems.

23 One improvement over the binary treatment approach to estimating the treatment effect is when using an event study, the variance of treatment issues are moot.

24 But in selecting only the never-treated as controls, the approach may have limited value for those situations where the number of units in the never-treated pool is extremely small.