Microeconometrics: Methods of Causal Inference

(Master in Economics)

Damian Clarke[1]

Semestre Primavera 2021

Last updated October 26, 2021

Universidad de Chile

[1] Associate Professor, University of Chile. Email dclarke@fen.uchile.cl. These notes have been brought together from a large range of sources, including a quantitive methods cause I taught at University of Oxford based on a much earlier version of these materials. A number of sections of these notes owe a great deal to earlier lecturers of that course, particularly to Andrew Zeitlen who taught this course over a number of years, and to Clément Imbert. The original notes from Andrew's classes are available on his website, and also in "Empirical Development Economics" (2014).

Background

We will use these notes as a guide to what will be covered in the Microeconometrics coures in the Master of Economics at the University of Chile. We will work through the notes in class and undertake a series of exercises on computer to examine various techniques. These notes and class discussion should act to guide your study for the end of year exam.

Along with each section of the notes, a list of suggested and required reading is provided. Required reading should act as a complement to your study of these notes; feel free to choose the reference which you prefer from the list of required readings where two options are listed. I will point you to any particularly relevant sections in class if it is only present in one of these. You are not expected to read all references listed in suggested readings. These are chosen as an illustration of the concepts taught and how these methods are actually used in the applied economics literature. At various points of the term you will be expected to give a brief presentation discussing a paper chosen from the suggested reading list, or other papers which you would like to propose (subject to confirmation with the professor). Readings like this can also be extremely useful as you move ahead with your own research, and in eventually writing up your thesis.

Contents

Chapter 1

Econometrics in Parallel Universes

1.1 An Introduction to Treatment Effects and the Potential Outcome Framework

Required Readings
Imbens and Wooldridge (2009): Sections 1-3.1 and 5.1
Angrist and Pischke (2009): Chapters 1-2

The treatment effects literature focuses on how to *causally interpret* the effect of some intervention (or treatment) on subsequent outcomes.

The use of treatment effects methods is frequent—in the academic literature as well as in the work of government and international organisations. Famous examples in the economics literature include—among many others—the effect of deworming medication on children's cognitive outcomes, the effect of having been involved in war on labour market earnings, the effect of microfinance receipt on small business profit, and the effect of certain types of political leaders on outcomes in their constituencies. The nature of the type of interventions examined using treatment effect methodologies is very broad. They may be interventions designed explicitly by researchers (such as those which are common in organisations like JPAL), they may be public policies such as anti-poverty programs, they may be environmentally imposed, such as exposure to pollution, or they may be a mixture of these, such as the PROGRESA/Oportunidades program which is an experimentally defined public policy. However, what all treatment effects methods have in common, regardless of the nature of the intervention, is a clear focus on identifying causal "treatment effects" by comparing a treated individual to an appropriately defined control individual.[1]

[1] Without loss of generality, you could replace "individual" with "firm" or some other unit of treatment. For

This may sound slightly different to what you have considered in your studies of econometrics so far. In previous econometrics courses, the consistent estimation of parameters of interest has relied upon assumptions regarding individual-level unobservables u_i, and their relationship (or lack thereof) with other variables of interest x_i. In this course however, estimation will be explicitly based on considering who is the appropriate *counterfactual* to be compared to the treated individual. Fortunately, while the way of thinking about these methods is different to what you have likely seen so far, many of the tools and assumptions that we make will have a very natural feel to you from earlier courses. We will once again encounter regressions, instrumental variables, and panel data at various points in this course, however the framework will generally explicitly refer to treatment effects based off counterfactual comparisons.

1.1.1 The Case for Parallel Universes

In the simplest sense, what treatment effects methods boil down to is the application of a 'parallel universe' thought experiment. In order to determine the effect that receipt of treatment has on a person, what we would really like to observe is precisely the same individual who lives their life in two nearly identical cases. In one universe, we would like to see what happens to the individual when they receive the treatment of interest, and in the other universe, we'd like to see the same individual in the same context, subject to the minor difference that they did not receive treatment. Then, without any complicated econometrics, we could infer that the causal impact of treatment is simply the difference between the individual's outcomes in these two worlds.[2]

In slightly more formal terms, we can think of an individual i, with observed characteristics x_i, assigned to treatment $w \in \{0, 1\}$, and with observed outcome y_i. In reality of course, we cannot run our thought experiment, as we observe only one of the two cases: either the individual is treated, in which case $w = 1$, or is untreated, with $w = 0$. The job for us as econometricians then is in answering the question: what would individual i have looked like if they had received treatment w' instead? (Or, in other words, what would have happened in the parallel universe?)

This question leads us to the Rubin Causal Model...

the sake simplicity, we will refer to the unit of treatment as "individuals" throughout the rest of these notes.

[2] This may seem very far fetched, but social scientists have expended a lot of effort in wriggling around the lack of an observed alternative universe. We could think, for example, of all the work on monozygotic twins as an—admittedly flawed—real world attempt at examining individuals with identical genetic material in parallel lives…

1.1.2 The Rubin Causal Model

The Rubin Causal Model (RCM) introduces a language that can be useful in clarifying thinking to answer that question. At first glance this way of modeling the question under study may seem very different from what you have seen so far in econometrics. In Section 1.1.3 of these notes we will return and relate this back to the kinds of empirical models with which you are already familiar. The RCM divides the evaluation problem into two distinct parts: a set of potential outcomes for each unit observed, and an assignment mechanism that assigns each unit to one and only one treatment at each point in time. We will examine these in turn.

Potential Outcomes

Let W_i be a random variable for each individual i that takes a value of 1 if they receive a particular treatment, and 0 otherwise.[3] We will be interested in a measurable outcome, Y.

For example, we may be interested in the impact of attending secondary school on subsequent labor-market earnings. In that case, w_i would take a value of unity only for those individuals who attend secondary school, and y would be a measure of their earnings. Examples of such analysis abound, and have even come to dominate much of the applied, microeconomic work in development. If you open up a recent issue of AEJ Applied Economics or AEJ Economic Policy, you will likely find many interesting examples of problems cast in this way.

Any given individual could be associated with either treatment (in which case $w_i = 1$) or its absence ($w_i = 0$). The RCM defines a pair of potential outcomes, (y_{1i}, y_{0i}) to these counterfactual states. So far, so good. However, there is a problem…At any point in time, only one of these potential outcomes will actually be observed, depending on the condition met in the following assignment mechanism:

$$y_i = \begin{cases} y_{1i}, & \text{if } w_i = 1 \\ y_{0i}, & \text{if } w_i = 0. \end{cases} \tag{1.1}$$

At this point it is worth explicitly making note that both of these outcomes together will never exist for a given i. If we observe y_{1i} (an individual's outcome under treatment) this precludes us from observing y_{0i}. Conversely, observing an individual's outcome in the absence of treatment implies that we will never observe the same unit under treatment. This is what Holland (1986) calls the "fundamental problem of causal inference": for the individuals who we observe under treatment we have to form an estimate of what they would have looked like if they had not been treated.

[3] In fact it is not necessary—and can be misleading—to think of the alternative to particular treatment as the *absence* of any intervention. Often we will be interested in comparing outcomes under two alternative treatments.

The observed outcome can therefore be written in terms of the outcome in the absence of treatment, plus the interaction between the treatment effect for that individual and the treatment dummy:

$$y_i = y_{0i} + (y_{1i} - y_{0i})w_i. \tag{1.2}$$

(Imbens and Wooldridge, 2009, pp. 10-11) provide a useful discussion of the advantages of thinking in terms of potential outcomes. Worth highlighting among these are:

1. The RCM forces the analyst to think of the causal effects of specific manipulations. Questions of the 'effect' of fixed individual characteristics (such as gender or race) sit less well here, or need to be carefully construed. A hard-line view is expressed by Holland (and Rubin): "NO CAUSATION WITHOUT MANIPULATION" (Holland (1986), emphasis original).

2. The RCM clarifies sources of uncertainty in estimating treatment effects. Uncertainty, in this case, is *not* simply a question of sampling variation. Access to the entire population of observed outcomes, y, would not redress the fact that only one potential outcome is observed for each individual unit, and so the counterfactual outcome must still be estimated—with some uncertainty—in such cases.

The Assignment mechanism

The second component of the data-generating process in the RCM is an assignment mechanism. The assignment mechanism describes the likelihood of receiving treatment, as a function of potential outcomes and observed covariates.

Assignment mechanisms can be features of an experimental design: notably, individuals could be randomly assigned to one treatment or another. Alternatively the assignment mechanism may be an economic or political decision-making process. We sometimes have a mixture of the two; for example, when we have a randomized controlled trial with imperfect compliance (which will be discussed much more in section 3.1 later in this lecture series).

Thinking in terms of potential outcomes and an assignment mechanism is immediately helpful in understanding when it is (and is not) appropriate to simply compare observed outcomes among the treated and observed outcomes among the untreated as a measure of the causal effects of a program/treatment. Note (Angrist and Pischke, 2009, p. 22) that

$$\begin{aligned}\underbrace{E[Y_i|W_i=1]-E[Y_i|W_i=0]}_{\text{Observed difference in average outcomes}} &= \underbrace{E[Y_{1i}|W_i=1]-E[Y_{0i}|W_i=1]}_{\text{average treatment effect on the treated}} \\ &+ \underbrace{E[Y_{0i}|W_i=1]-E[Y_{0i}|W_i=0]}_{\text{selection bias}},\end{aligned} \tag{1.3}$$

by simply adding and subtracting the term in the middle (note that these two terms are the same!).

This is quite an elegant formula, and a very elegant idea. If we consider each of the terms on the right-hand side of equation 1.3, first:

$$E[Y_{1i}|W_i = 1] - E[Y_{0i}|W_i = 1].$$

This is our estimand of interest, and is the average causal effect of treatment on those who received treatment. This term is capturing the average difference between what actually happens to the treated when they were treated ($E[Y_{1i}|W_i = 1]$), and what would have happened to the treated had they not been treated ($E[Y_{0i}|W_i = 1]$).

The second term refers to the bias potentially inherent in the assignment mechanism:

$$E[Y_{0i}|W_i = 1] - E[Y_{0i}|W_i = 0].$$

What would have happened to the treated had they not been treated (once again, $E[Y_{0i}|W_i = 1]$), may be quite different to what actually happened to the untreated group in practice ($E[Y_{0i}|W_i = 0]$). It is worth asking yourself at this point if this all makes sense to you. In the above outcomes, what do we (as econometricians) see? What don't we see? What sort of assumptions will we need to make if we want to infer causality based only on observable outcomes? We will return to discuss these assumptions in more depth soon.

As we will see, when potential outcomes are uncorrelated with treatment status—as is the case in a randomized trial with perfect compliance—then the selection bias term in equation 1.3 is equal to zero. Due to randomisation, the treated and control individuals should look no different on average, and as such, their potential outcomes in each case should be identical. In this ideal set-up, comparison of means by treatment status then gives the treatment effect experienced by those who received the treatment.

In general, the assignment of an individual to treatment status w_i may depend on observable characteristics, x_i. It may also depend on unobserved determinants of the potential outcomes. In this way we can, in general, have

$$w_i = f(x_i, y_{1i}, y_{0i}). \tag{1.4}$$

This is very broad, stating that assignment can depend upon observable characteristics (generally not a problem), but also could depend upon the potential outcomes themselves (which will, in general, require attention).[4] As we will see in the remainder of this course, the appro-

[4] As a simple example, we could consider the example of a program where the individuals who choose to enter are those who would do the worst without the program. Using non-treated individuals as a counterfactual in this case is clearly not appropriate, as their experience without the program is better than what would be expected were

priateness of alternative estimators will hinge crucially on whether we are willing to assume that selection is a function only of observable characteristics, or whether we want to allow it to depend on unobservable characteristics as well.

Estimands of Interest

In this general framework, we have not assumed that potential outcomes (Y_{0i}, Y_{1i}) are the same across all individuals, or even that the *difference* between the potential outcomes is constant across individuals. This permits alternative definitions of program impact. For now we will focus on two:[5]

- **Average Treatment Effect (ATE)**: $E[Y_1 - Y_0]$
- **Average Treatment Effect on the Treated (ATT)**: $E[Y_1 - Y_0|W = 1]$

The first of these, the ATE, represents the average improvement that would be experienced by all members of the population under study, if they were all treated. The ATT, on the other hand, is the average treatment effect *actually experienced* in the sub-population of those who received treatment. Depending on the use of our econometrics, the statistic we will be interested in will vary. For example, if we are interested in assessing the impact of a targeted anti-poverty program, it seems unlikely that we would be interested in the ATE in the whole population, many of whom are not eligible for the program, and would likely prefer the ATT. On the other hand, if we were aiming to assess the impact of a program that is planned to roll-out to the whole population over time, the ATE is precisely what we would like to know.

We will sometimes (and throughout the remainder of this section) assume that treatment effects are *homogeneous*; i.e., that they are the same throughout the population. In this case, clearly, the ATT and ATE will be the same. The two measures of program impact will diverge, however, when there is *heterogeneity in treatment response* (or potential outcomes) across individuals, and when selection into treatment—the assignment mechanism—is not independent of these potential outcomes.

To see why the ATT and ATE will often not be the same, consider analyzing the effect of obtaining secondary schooling on subsequent income. The returns to secondary schooling will vary by individual: those with greater natural ability or connections in the employment market may be better placed to benefit from additional schooling. If it is also the case that those who end up receiving schooling are those with higher returns, then the ATT will be greater than the

the treatment group not to participate.

[5] In the following lecture, we will discuss non-compliance in more detail. We will then introduce a third measure, the Intent-to-Treat (ITT) effect.

ATE. Such concerns are central to the '*scaling up*' of development interventions: if the ATT and the ATE differ, then intervening to obtain complete coverage may not yield the expected results.

1.1.3 Returning to Regressions

Thus far, the language of treatment effects may seem a bit foreign to the regression framework to which you have become accustomed. This need not be so. In fact, starting from a slightly more general version of the potential outcomes framework can help to clarify the assumptions underlying regressions used for causal inference.

Let's begin by assuming that there are no covariates—just the observed outcome, Y, and a treatment indicator, W. It will be helpful to write μ_0, μ_1 as the population means of the potential outcomes Y_0, Y_1 respectively. These values are generally our estimands of interest, and can be compared to the coefficients you have been estimating in regession models throughout the whole course. Let e_{0i}, e_{1i} be a mean-zero, individual-specific error term, so that we can write:

$$y_{0i} = \mu_0 + e_{0i} \tag{1.5}$$

$$y_{1i} = \mu_1 + e_{1i}. \tag{1.6}$$

Then, recalling equation (1.2), we can write the observed outcome as

$$y_i = \mu_0 + \underbrace{(\mu_1 - \mu_0)}_{\tau} w_i + \underbrace{e_{0i} + (e_{1i} - e_{0i})w_i}_{e_i}. \tag{1.7}$$

Thus we can see that a regression of y on w will produce a consistent estimate of the average treatment effect only if w is uncorrelated with the compound error term, e_i. This holds when treatment assignment is uncorrelated with potential outcomes—an assumption that we will introduce in Section 1.1.4 as *unconfoundedness*.

Covariates can also be accommodated in this framework. Consider a covariate X_i. For ease of exposition define $\bar{x}$ as the population average of x; we can then write:

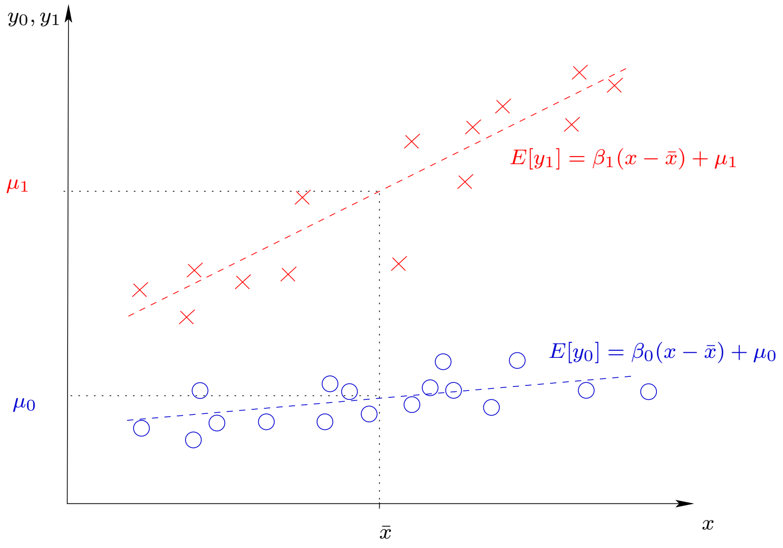
$$y_{0i} = \mu_0 + \beta_0(x_i - \bar{x}) + e_{0i} \tag{1.8}$$

$$y_{1i} = \mu_1 + \beta_1(x_i - \bar{x}) + e_{1i}. \tag{1.9}$$

Notice here that we can allow the coefficients, β, to vary according to treatment status. This is illustrated in Figure 1.1.

The ATE is still given by $\mu_1 - \mu_0$, and we can still include x as a regressor (the reasons for doing so are discussed in the next section). But we may now want to take explicit care to

Figure 1.1: Treatment effect heterogeneity with observable characteristic x



let the relationship between x and y depend on treatment status, and to incorporate this into our estimates of the treatment effect. This allows us to flexibly model the situation in which $\beta_0 \neq \beta_1$ in equations 1.8 and 1.9. There are many real-life examples where this might be the case: for example, the effect of social networks on earnings might be stronger among those with secondary education (a treatment of interest) than among those without. We will return to a more extensive discussion of heterogeneity in the lectures which follow, and particularly, section 3.1 of these notes.

Let us leave aside—for the moment—the issue of varying coefficients. The key question then becomes, under what circumstances will a regression of the form above give consistent estimates of the effect of treatment W? We now turn to this.

1.1.4 Identification

The simplest case in the analysis of treatment effects occurs when the following three assumptions hold.

Assumption 1. *Stable Unit Treatment Value Assumption (SUTVA).*

Potential outcomes Y_{0i}, Y_{1i} are independent of W_j, $\forall j \neq i$.

This is the assumption that the treatment received by one unit does not affect the potential outcomes of another—that is, that there are no externalities from treatment. When SUTVA fails, the typical responses are either to change the unit of randomization/analysis, so as to internalize the externality; or to estimate the externalities direcly. See in particular Miguel and

Kremer (2004) for a paper that grapples with such externalities[6]. However, we will maintain the SUTVA assumption throughout this and the next lecture, unless otherwise specified.

While not explicitly built into SUTVA, the importance of effects and one's *own* treatment status is something that we will want to think carefully about when considering the scope of results. Both John Henry Effects and Hawthorne Effects will lead to a situation where we may assign to the treatment an effect which is actually due to people realising that they are participating in a trial.

Assumption 2. *Unconfoundedness*

$$(Y_{0i}, Y_{1i}) \perp\!\!\!\perp W_i | X_i$$

Conditional on covariates X_i, W is independent of potential outcomes. Variations of this assumption are also known as conditional mean independence *and* selection on observables*.*

As suggested by equation (1.7), unconfoundedness is required for simple regression to yield an unbiased estimate of the ATT, τ. This is also evident in the decomposition of equation (1.3): unconfoundedness ensures that $E[Y_{0i}|W_i = 1] = E[Y_{0i}|W_i = 0]$. We may not always be confident that unconfoundedness holds *unconditionally*, but in some cases conditioning on a set of characteristics X can strengthen the case for the applicability of this assumption.

It is important to note that this is a particularly strong assumption. If we are willing to make an assumption of this type, it buys us identification under a very wide range of settings. However, we should always ask ourselves whether we believe the assumption in each circumstance in which we call upon it. This assumption is not dissimilar, in magnitude or scope, to the exogeneity assumption from the Gauss-Markov theorem that has been present in earlier econometrics courses.

Assumption 3. *Overlap*

$$0 < \text{Pr}[W_i = 1|X_i] < 1$$

The assumption of overlap implies that, across the support of X, we observe both treated and untreated individuals. In other words, for every combination of X_i, at least one treated and one untreated individual exists. Note this is an assumption about the *population* rather than about the sample; the hazards of random sampling make it highly likely (especially in the case of multiple and discrete regressors) that we will not observe both treated and untreated individuals with exactly the same value of these covariates.

[6] These questions are far from trivial. You may be familiar with the challenges and critiques which arose during the so-called "Worm Wars" (see for example Davey et al. (2015); Hicks et al. (2015)). This was an example where the precise issues which we are discussing in these four lectures (the consistent estimate of treatment effects) spilled over into the popular press.

Assumptions 2 and 3 are sometimes known together as the condition of "strongly ignorable treatment assignment" (Rosenbaum and Rubin, 1983). The identification of a conditional average treatment effect $\tau(x)$ under unconfoundedness and overlap can be shown as follows:

$$\begin{aligned} \tau(x) &= E[Y_{1i} - Y_{0i}|X_i = x] && (1.10) \\ &= E[Y_{1i}|X_i = x] - E[Y_{0i}|X_i = x] && (1.11) \\ &= E[Y_{1i}|X_i = x, W_i = 1] - E[Y_{0i}|X_i = x, W_i = 0] && (1.12) \\ &= E[Y|X_i = x, W_i = 1] - E[Y|X_i = x, W_i = 0] && (1.13) \end{aligned}$$

Equation (1.10) is given by the definition of the average treatment effect. Equation (1.11) follows from the linearity of the (conditional) expectations operator. Unconfoundedness is used to justify the move to equation (1.12): the potential outcome under treatment is the same in the treated group as it is for the population as a whole, for given covariates x, and likewise for the potential outcome under control. Equation (1.13) highlights that these quantities can be observed by population averages.

Equation 1.12 is central for us. This is the first time that we are actually able to say something using values observed in the real world rather than simply using theoretical potential outcomes (or in other words, we now have an identified parameter). This makes explicit the importance of the unconfoundedness assumption for identification in this context.

1.2 Constructing a Counterfactual with Observables

Required Readings
Imbens and Wooldridge (2009): Sections 4 and 5 (Don't worry about 5.2 and 5.9)
Angrist and Pischke (2009): Sections 3.2 and 3.3

Suggested Readings
Dehejia and Wahba (2002)
Diaz and Handa (2006)
Jensen (2010)
Banerjee and Duflo (2009)

This section could alternatively be called "estimation under unconfoundedness". Once we make assumptions of (conditional or unconditional) unconfoundedness, we have a range of estimation methods at our disposal. As unconfoundedness solves the business of the assignment mechanism by making it completely observable, all we have left is to recover estimates of these treatment effects by using data. This is now a technical issue, which we turn to here.

1.2.1 Unconditional unconfoundedness: Comparison of Means

The simplest case occurs when $(Y_1, Y_0) \perp\!\!\!\perp W$, without conditioning on any covariates. Where this assumption holds, we need only compare means in the treated and untreated groups, as already shown. The ATE can be estimated by a **difference-in-means** estimator of the form:

$$\hat{\tau} = \sum_{i=1}^{N_1} \lambda_i Y_i - \sum_{i=1}^{N_0} \lambda_i Y_i, \tag{1.14}$$

where N_0, N_1 are the number of treated and untreated individuals in the sample, respectively, and where the weights in each group add up to one:

$$\sum_{i:W_i=1} \lambda_i = 1$$
$$\sum_{i:W_i=0} \lambda_i = 1.$$

A straightforward way to implement this in Stata or your favourite computer language for econometrics is just to regress outcome y on a dummy variable for treatment status.

When will unconditional unconfoundedness hold? It is likely only to hold *globally* (that is, for the entire population under study) in the case of a randomized controlled trial with perfect compliance. This is the reason that claims are sometimes made that such experiments provide a 'gold standard' in program evaluation. Since the regression can be performed without controls, it may be less susceptible to data mining and other forms of manipulation by the researcher, a point we turn to in the final section of these notes.

Even in a RCT however, there are a number of important considerations, especially when putting this into practice. Issues such as *how* to randomise (is it okay to just flip a coin, for example?), testing for balance of covariates between treatment and control groups, the use of stratified or blocked randomisation, and power calculations are all things that come up in this context. We won't go in to too great depth here, however if you ever find yourself working in a situation where you are participating in an RCT, an excellent place to start is by reading Glennerster and Takavarasha's 2013 "Running Randomized Evaluations: A Practical Guide", a comprehensive applied manual with an accompanying webpage: http://runningres.com/.

The handbook chapter of Duflo et al. (2007) also provides an extremely useful overview, particularly focused on development economics. This also provides hands-on discussion of the practicalities involved in implementing randomized control trials along with some key considerations such as details related to working with partners for implementation, the procedure of piloting projects, different methods of randomization elements related to sampling and sample size, and data collection. We will discuss some of these in more length later, particularly in

Chapter 4 when we discuss power in hypothesis tests. Each of these considerations has many 'moving parts' and is worth reading in full. For example, when considering the *way* in which randomization can take place, Duflo et al. (2007) list (i) the oversubscription method, where participants can be chosen randomly from applicants where more applicants than spots exist, (ii) Randomized order of phase-in, where all individuals eventually receive treatment, but the timing is staggered, (iii) within group randomization, where certain sub-groups in each group receive treatment, or (iv) encouragement designs, where rather than randomizing the program itself, researchers provide random groups encouragement of some sort to participate in a program.

While RCTs allow us to quite credibly make the unconfoundedness assumption, such trials are not easy to implement and will not be able to answer all questions—an issue to which we return to extensively in the all the lectures which follow. Deaton (2009) provides a critique. For now we may note that:

- Randomized controls are expensive and time-intensive to run;
- The set of questions that can be investigated with randomized experiments is a strict subset of the set of interesting questions in economics;
- Evidence from RCTs is subject to the same problems when it comes to extrapolating out of the sample under study as is evidence from other study designs.
- Attrition and selection into/out of treatment and control groups pose serious challenges for estimation.

This is something followed up in Deaton (2020), which pays particular attention to important *ethical* considerations behind RCTs in economics. This is not a trivial concern, and something of central importance in research, and this paper is well-worth reading.

While experiments do very well in terms of *internal validity*—they identify the treatment effect for some subpopulation within the sample—they are no guarantee of *external validity*. Replication (which may provide evidence that treatment effects are homogeneous, or vary in predictable ways with measurable characteristics) and, ultimately, theory, are required.

Unconfoundedness will hold globally by design in RCTs. In a less controlled (by the econometrician) setting, we may be willing to assume that unconditional unconfoundedness holds *locally* in some region. This is the basis for regression discontinuity design, to be discussed later in the lecture series (section 3.2).

1.2.2 Regressions

Absent a RCT, unconfoundedness is unlikely to hold unconditionally. In nearly all other cases in which we will be interested, there will be some reason why individuals receive treatment – be it an explicitly targeted program, or individuals choosing to participate in a program given the incentives they face. As a start, we may be able to make the unconfoundedness assumption less stringent by conditioning on a set of characteristics, X. By now the most familiar way of doing so is through multivariate regression. If we are able to perfectly measure the characteristics that are correlated with both potential outcomes and the assignment mechanism, then this problem can be resolved with regression.

Recall the potential outcomes framework with covariates, from equations (1.8) and (1.9). Let's combine these seperate equations into one regression model, where we assume a linear functional form for the relationship between x and each of the potential outcomes (note that this need not be the case). This leads to a regression of the form:

$$y_i = \mu_0 + (\mu_1 - \mu_0)w_i + \beta_0(x_i - \bar{x}) + (\beta_1 - \beta_0)(x_i - \bar{x})w_i + e_{0i} + (e_{1i} - e_{0i})w_i. \quad (1.15)$$

Often it is assumed that $\beta_0 = \beta_1 = \beta$, in which case this expression simplifies to:

$$y_i = \mu_0 + (\mu_1 - \mu_0)w_i + \beta(x_i - \bar{x}) + e_{0i} + (e_{1i} - e_{0i})w_i. \quad (1.16)$$

Under (conditional) unconfoundedness, $E[e_{0i} + (e_{1i} - e_{0i})w_i | X_i] = 0$, so the unobservable does not create bias in the regression.

But this foreshadows the importance of *either* getting the functional form for β exactly right, or else having the x characteristics balanced across treatment and control groups. If covariates are not balanced, then omission of the term $(\beta_1 - \beta_0)(x_i - \bar{x})w_i$ introduces a correlation between w and the error term, biasing estimates of the ATE.

It may be tempting to conclude that it is best to err on the side of including covariates X. And indeed, in many cases this will be the case. You have likely observed in earlier econometrics courses that including irrelevant covariates in a regression does not bias coefficients, while the omission of relevant covariates generally does. However there is an important class of covariates that should be omitted from a regression approach: intermediate outcomes.

The logic here is simple. Suppose the treatment of interest, W affects a second variable, so that $E[X|W] = \delta W$, and that both X and W have direct effects on the outcome of interest Y. In this case, if we are interested in the impact of W on Y, we want a total derivative—inclusive of the effect that operates through intermediate outcome X. Conditioning on X in a regression would in this case bias (towards 0) such an estimate.

As Angrist and Pischke (2009) point out, such intermediate outcomes may depend both on unobserved factors that we would like to 'purge' from their potential confounding influence on the estimates, as well as a causal effect stemming from W. In this case, the researcher faces a trade-off between two sources of bias.

As an example, imagine if we were interested in following up the well known Miguel and Kremer (2004) worms trial to look at the effect of deworming drugs on eventual labour market outcomes of recipients (see for example Baird et al. (2016)). We would quite quickly reach the question of whether we should include education as a control. Education has large returns on the labour market, and seems like a relevant control in a labour market returns regression. But, at the same time, any difference in education between treatment and control may be largely due to the effect of treatment (deworming) itself. The way we would decide to move forward is not entirely clear, and would require careful consideration of what inclusion or exclusion of the controls would imply for our parameter estimates.

1.2.3 Probability of Treatment, Propensity Score, and Matching

Unconfoundedness, when combined with regression, gives consistent estimates of the ATT. But we have seen that, when conditioning on a vector of covariates X is required for this assumption to hold, results may be sensitive to functional form. One response is to use very flexible functional forms in X, but given the degrees of freedom requirements this is not always practical or ideal. A common family of alternatives to regressions of the sort described in Section 1.2.2 are based on the propensity score.

Begin by defining the *propensity score*, $p(x) = \Pr[W = 1|X = x]$, as the probability of being treated, conditional on characteristics x. Propensity score methods are based on the observation that, once we assume unconfoundedness, the treatment indicator and potential outcomes will be independent of one another conditional on the propensity score Rosenbaum and Rubin (1983).

Theorem 1. *Propensity score theorem*

Suppose unconfoundedness holds, such that $W_i \perp\!\!\!\perp (Y_{0i}, Y_{1i})|X_i$, and define the propensity score as above. Then potential outcomes are independent of the assignment mechanism conditional only on the propensity score: $W_i \perp\!\!\!\perp (Y_{0i}, Y_{1i})|p(X_i)$.

The intuition for this result comes from the observation that *even without unconfoundedness*, $W_i \perp\!\!\!\perp X_i|p(X_i)$. See Angrist and Pischke (2009) for a useful discussion. In a general sense, as the propensity score is capturing the assignment mechanism, conditional on the propensity score, all that remains of the Rubin Causal Model is the difference in potential outcomes between treated and untreated individuals.

Having established that we need only condition on the propensity score in order to ensure independence of the assignment mechanism and the potential outcomes, we have a range of estimating techniques available.

Regression using the propensity score

Possibly the most straightforward use of the propensity score is to use it to augment a simple regression of observed outcomes on treatment status. In practice this entails first estimating the propensity score (typically with a logit or probit),[7]and then including this generated regressor in a regression of the form:

$$y_i = \tau w_i + \phi \widehat{p(x_i)} + e_i \tag{1.17}$$

If the relationship between the propensity score and potential outcomes is in fact a linear one, then the inclusion of $p(X)$ purges this regression of any contamination between the treatment status w and the error term (recall that the error term contains the individual-specific variation around the population means of the potential outcomes).

At first glance, this seems to offer a pair of benefits—but these are not straightforward.

First, regression using the propensity score seems to be a solution for a degrees of freedom problem, in that it is no longer necessary to control for a (potentially high dimension) X in the regression on potential outcomes. However, this is not the case, since p is a function of the full set of covariates. This is most easily seen when the propensity score is estimated by a linear probability model, in which case the estimates are *exactly the same* as those obtained by inclusion of X directly.

Second, regression using the propensity score seems to allow us to be agnostic about the functional form relating X to potential outcomes Y_{0i}, Y_{1i}. Often these functional forms have been the subject of long debates (for example, in the case of agricultural production functions or earnings functions), whereas our interest here is simply in the use of X to partial out any correlation between the assignment mechanism for W and the potential outcomes. However, regression using the propensity score as in equation (1.17) requires us to correctly specify the relationship between the propensity score and the potential outcomes, an object for which theory and accumulated evidence provide even less of a guide, while at the same time requiring us to correctly specify the function $p(X)$. This is partly solved by including higher-order polynomial functions of p, but at the expense of the parsimony that is the chief advantage of this approach. The two estimates discussed next—weighting and matching using the propensity

[7] In Stata, propensity scores can be estimated using the `pscore` command. Alternatively logit, probit (or for that matter linear probability) models can be combined with the `predict` post-estimation command to generate the propensity scores for each observed unit. As of version 13 of Stata, there is a new series of commands contained in the `teffects` library which includes a propensity score module `pscore`.

score—have the advantage of allowing us to be truly agnostic about the relationship between potential outcomes and $p(X)$.

As a final precaution in the case that you wish to combine a propensity score estimate with regression methods, it is important to note that in such an approach (as with instrumental variables estimates when done 'by hand'), standard errors must be corrected for the presence of generated regressors. Bootstrap or other resampling methods are often the easiest route of calculating standard errors in circumstances such as these.

Weighting by the propensity score

Under unconfoundedness, the propensity score can be used to construct weights that provide consistent estimates of the ATE. This approach is based on the observation that (again, under unconfoundedness)

$$E[Y_{1i}] = E\left[\frac{Y_i W_i}{p(X_i)}\right] \tag{1.18}$$

and

$$E[Y_{0i}] = E\left[\frac{Y_i(1-W_i)}{(1-p(X_i))}\right]. \tag{1.19}$$

To see why, note that, as discussed in Angrist and Pischke (2009, p. 82), equation 1.18 can be shown to hold as follows:

$$\begin{aligned} E\left[\frac{Y_i W_i}{p(X_i)}\right] &= E\left\{E\left[\frac{Y_i W_i}{p(X_i)}\right]\Bigg| X_i\right\} \\ &= \frac{E[Y_i|W_i=1,X_i]p(X_i)}{p(X_i)} \\ &= E[Y_{1i}|W_i=1,X_i] = E[Y_{1i}|X_i] \end{aligned}$$

and a similar process can be followed for $E[Y_{0i}]$ (equation 1.19). Combining these gives an estimate of the ATE:

$$\begin{aligned} E[Y_{1i}-Y_{0i}] &= E\left[\frac{Y_i W_i}{p(X_i)} - \frac{Y_i(1-W_i)}{(1-p(X_i))}\right] \\ &= E\left[\frac{(W_i-p(X_i))Y_i}{p(X_i)(1-p(X_i))}\right] \end{aligned} \tag{1.20}$$

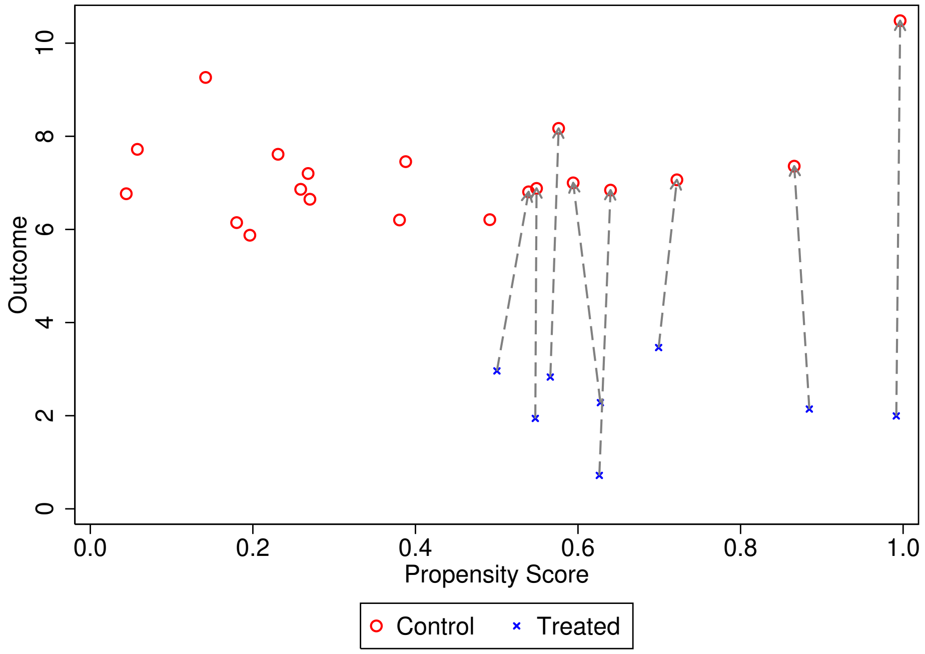
which can be estimated using sample estimates of $p(X)$. This idea can be thought of as framing the problem of analyzing treatment effects as one of non-random sampling. Although this insight allows us to avoid making functional form assumptions about the relationship between potential outcomes and X, it does require a consistent estimate of the propensity score.

Matching on the propensity score

An alternative and perhaps more intuitive set of estimators are based on matching. To begin, note that under Assumption 3, in a large enough sample it should be possible to match treated observations with untreated observations that share the same value of the covariate vector X. When the covariates are discrete variables, this amounts to ensuring that we have both treated and untreated observations in all the 'bins' spanned by the support of X. However, in finite samples and in particular with many, continuous regressors in X, exact matching becomes problematic: we suffer from a curse of dimensionality.

Application of the propensity score theorem tells us that it is sufficient to match on the basis of $p(X)$, rather than matching on the full covariate vector X.

Figure 1.2: Propensity-score matching using nearest-neighbor matching



Once we have established that our data—or a subset of observations—satisfy the requirements of common support and conditional mean independence, we can obtain an estimate of the ATT by:

$$ATT^M = \frac{1}{N_T} \sum_{i:w_i=1} \left(y_{1,i} - \sum_{j:w_j=0} \phi(i,j) y_{0,j} \right) \tag{1.21}$$

where $\{w = 1\}$ is the set of treated individuals, $\{w = 0\}$ is the set of untreated individuals, and $\phi(i, j)$ is a weight assigned to each untreated individual—which will depend on the particular matching method. Notice that $\sum_{j:w_j=0} \phi(i, j) y_{0,j}$ is our estimate of the counterfactual outcome for treated individual j.

The issue now is how to calculate the weight. There are several possibilities. Two common approaches include:

- Nearest-neighbor matching: find, for each treated individual, the untreated individual with the most similar propensity score. $\phi(i,j) = 1$ for that j, and $\phi(i,k) = 0$ for all others.

- Kernel matching: Let the weights be a function of the "distance" between i and j, with the most weight put on observations that are close to one another, and decreasing weight for observations farther away.

Alternative matching methods also exist, including minimizing the Mahalanobis distance and optimising both the neighbours to be used and their weights together in a single optimisation problem. The Mahalanobis matching procedure seeks to directly minimize a single measure of distance based on the imbalance in covariates X. Consider two observations i and j with vectors of observable characteristics X_i and X_j respectively. The Mahalanobis metric is to calculate their "distance" as:

$$M(X_i, X_j) = \sqrt{(X_i - X_j)'S^{-1}(X_i - X_j)}$$

where S refers to the sample covariance matrix of X. A "match" can then be sought based on the units which are closest in these measures. Note that alternative matching methods can give very different answers—we will see this ourselves in the data exercise. A limitation of propensity-score approaches is that there is relatively little formal guidance as to the appropriate choice of matching method. Relatively recent work from King and Nielsen (2019) points to additional concerns, specifically with propensity score matching, with both losses in efficiency due to the removal of observations, and at times even increases in bias. In general, all told, this suggests that propensity score matching should be avoided as a technique for causal analysis.

Matching methods (including propensity scores) can be combined with difference in differences (DiD) techniques. As in Gilligan and Hoddinot (2007), we could estimate:

$$ATT^{DIDM} = \frac{1}{N_T} \sum_{i \in \{w=1\}} \left(y_{1,i,t} - y_{1,i,t-1} - \sum_{j \in \{w=0\}} \phi(i,j)(y_{0,j} - y_{0,j,t-1}) \right) \tag{1.22}$$

which compares change in outcomes for treated individuals with a weighted sum of changes in outcomes for comparison individuals. We will return in far more detail to difference in differences methods in section 2 of these notes.

1.2.4 Matching methods versus regression

There is no general solution to the problem of whether (appropriately chosen) matching or regression methods should be preferred as ways of estimating treatment effects under conditionañ unconfoundedness—the appropriate answer will depend on the case. Of course, in general, if a combination of methods leads to conclusions which are broadly similar, this will give us much greater confidence in the validity of our estimates.

Advantages of propensity score/matching:

- Does not require functional form assumptions about the relationship between Y and the X covariates. As such it avoids problems of extrapolation: if the support of some X variables is very different across treated and untreated observations in the sample, then we will be forced to extrapolate the relationship between x and potential outcomes in order to estimate the treatment effect under regression (to see this, consider allowing the β to vary by treatment status).
- Can potentially resolve the 'curse of dimensionality' in matching problems.

Disadvantages

- Shifts the problem of functional form: must correctly specify $e(x) = \Pr[W = 1|X = x]$. Note that since most candidate estimates (probit, logit, etc) are relatively similar for probabilities near 1/2, these methods may be more appealing when there are few observations with very high or very low predicted probabilities of treatment.
- Matching on the basis of propensity score proves to be very sensitive to the particular matching method used.
- Asymptotic standard errors under propensity score matching are higher than under linear regression, even when we have the 'true' functional form—this is the price of agnosticism. In small samples, however, this may be less of an issue Angrist and Pischke (2009).

1.2.5 Some Points on Inference

In the case of regressions or comparison of mean estimators, typically inference—the procedure allowing for the construction of standard errors, confidence intervals, and eventually p-values—can be conducted using standard analytical formulae for variance. However, these are generally asymptotically valid, and based on strong assumptions such as normality of the residual terms. While these are still the most commonly used inference procedure, and increasingly common procedure consists of conducting “randomization inference”.

Randomization inference, originally laid out in Fisher (1925, 1935) provides an alternative means of inference which is valid in small samples, and can be conducted by simply taking the true data and randomly permuting (or shuffling) a treatment status throughout observations and then re-estimating the treatment effect, before counting how many times these randomly generated statistics exceed the original treatment value.

Randomization inference follows from the idea underlying “Fisher’s exact test” using contingency tables. Consider a case where we have 6 individuals, 3 of whom randomly receive a treatment, and 3 of whom randomly receive a placebo (control). If we compare the average between these two groups assuming balance, we can calculate the treatment effect as the difference in means. If we wish to formally test whether this is significantly different to zero, with 6 observations a t-test, and certainly something based upon asymptotic approximations, will likely not be appropriate. So to see whether this treatment effect is actually something that is significantly different to zero, one way to proceed would be to compare it to many samples where, in theory, no effect should exist, and ask how extreme the effect is compared to these samples. The logic behind the exact test, and randomization inference generally, is that we can generate such samples by randomly permuting the treatment status within the sample, holding all outcomes fixed, and simply considering the ‘treatment effects’ in all possible re-shuffled treatment cases. In order to calculate a p-value, we can ask how extreme the observed real treatment effect was compared with all the possible permuted treatment effects if the treatment status had been simply assigned at random to the same number of observations in this group.

It is perhaps useful to see a simple example. Consider the case of 6 units, with 3 observations randomly assigned treatment. Imagine that the observed outcomes were then, in the treatment group: $(34, 27, 29)$, and in the control group: $(14, 18, 24)$. A simple comparison of means estimator suggests that the treatment effect is 11.33. To calculate a p-value, we can permute all the possible combinations, and ask what proportion of these are greater than or equal to this treatment effect. If we consider random orderings of 6 units, this suggests that there are $6!$ possible combinations, but in reality, as we are randomly choosing 3 units from these 6 to assign a permuted treatment status, the actual value of different combinations is $\binom{6}{3} = \frac{6!}{3!*(6-3)!} = 20$. We document each of these possible permutations, as well as their permuted treatment effect in Table 1.1. In this case, we can see that only 1 of the 20 different permutations is greater than or equal to 11.33 (the original treatment status). Suggesting an exact p-value of $1/20 = 0.05$.

These methods are formally discussed in Athey and Imbens (2017) (among other places). Using their notation, the p-value we refer to above is denoted as:

$$p = pr(|T^{ave}(W, Y^{obs}, X)| \geq |T^{ave}(W^{obs}, Y^{obs}, X)|).$$

Table 1.1: A Simple Illustration of Randomization Inference

Permutation	Treatment			Control			Estimate
Original (1)	34	27	29	14	18	24	11.33
2	34	27	14	29	18	24	1.33
3	34	27	18	14	29	24	4
4	34	27	24	14	18	29	8
5	34	14	29	27	18	24	2.67
6	34	18	29	14	27	24	5.33
7	34	24	29	14	18	27	9.33
8	14	27	29	34	18	24	-2
9	18	27	29	14	34	24	0.67
10	24	27	29	14	18	34	4.67
11	34	14	18	27	29	24	-4.67
12	34	14	24	27	18	29	-0.67
13	34	18	24	14	27	29	2
14	14	27	18	34	29	24	-9.33
15	14	27	24	34	18	29	-5.33
16	18	27	24	14	34	29	-2.67
17	14	18	29	34	27	24	-8
18	14	24	29	34	18	27	-4
19	18	24	29	14	34	27	-1.33
20	14	18	24	34	27	29	-11.33

where T^{ave} refers to the statistic of interest (in our case the difference in means), and we can see that the left-hand side of this equation is the original treatment effect where each units true outcome Y is accompanied by its true treatment assignment W, whereas the right-hand side considers permutations where each Y is associated with randomly assigned treatment statuses W. One of the strengths of this randomization inference is that permutation can be performed over the same level of treatment assignment as in the original experiment, for example allowing for clustered treatments.

A nice very applied discussion is provided in the paper by Heß (2017), which introduces Stata tools to deal with randomization inference. It also shows the implementation for this in a particular paper, that of Fujiwara and Wantchekon (2013). Another recent paper implementing these methods is that of Baranov et al. (2020), who estimate the effect of random participation in a large psychotherapy program that reduced post-partum depression of mothers in Pakistan on their long-term well being, financial empowerment, and investments in children.

Finally, note that in general with a larger number of observations, we cannot calculate an exact test,[8] and thus generally we simply calculate a relatively large number of permuta-

[8] Even with quite moderate number of observations, the total number of possible combinations grows very quickly. For example, while 10 units with 5 randomly assigned treatments gives a reasonably manageable $\binom{10}{5}$ $= 252$ possible combinations, this grows quickly, with 20 units and 10 treatments resulting in $\binom{20}{10} = 184,756$

tions at random. Some very applied advice is given by the Development Impact Evaluation unit in the World Bank at the following page https://dimewiki.worldbank.org/wiki/Randomization_Inference, suggesting the following steps:

1. Preserve the original treatment assignment.
2. Generate placebo treatment statuses according to the original assignment method.
3. Estimate the original regression equation with an additional term for the placebo treatment.
4. Repeat #1–3.
5. The randomization inference p-value is the proportion of times the placebo treatment effect was larger than the estimated treatment effect.

possible combinations, and 30 units and 15 treatments a massive $\binom{30}{15}$=155,117,520 possible combinations.

Empirical Exercise 1: PROGRESA

Instructions: We will be using data from the conditional cash transfer program PROGRESA. This randomized treatment at the level of the community, where all people living below a poverty threshold received treatment in the treatment period if they lived in the treatment community, and all others did not receive treatment. For this, the dataset PROGRESA.dta is supplied. This dataset has observations on an individuals treatment (progresa), student enrollment (enrolled) the time period (t), whether the child lives in the treatment community (tcomm) and various other covariates. The data is a panel, with the children observed in two periods. The unique child identifier is called iid.

Please also note, that this assignment requires the use of two user written ado files. These are `psmatch2` and `pscore`. `pscore` is circulated with the Stata Journal, so cannot be installed using `ssc install`. To install both sets of ado files, the following commands should be used:

```
ssc install psmatch2
net from http://www.stata-journal.com/software/sj2-4
net install st0026
```

Questions:

(A) Descriptive Statistics Open that data and generate the following descriptive statistics to get a feel for the data:

1. How many children are there in the data? Is the panel strongly balanced?
2. What percent of children from the data live in treatment villages?
3. Is the program correctly targeted (ie, where only poor children treated)?
4. Did all poor children in treatment municipalities receive treatment?
5. The variable “score” is a poverty score. How does the poverty score look for poor and non-poor individuals?

(B) Experimental Evidence of the Impact We will now examine the experimental outcomes of PROGRESA. In this section, we will thus focus on period 2 only (the period in which the experiment was conducted).

1. What is the comparison of means estimator of the effect of PROGRESA among eligible children in the period of the experiment when considering the outcome of interest “enrolled”?
2. What are the assumptions which must hold for this to be an unbiased estimate?

3. Does this seem reasonable in the context?

(C) Non-experimental analysis: Difference-in-difference Suppose that PROGRESA had not conducted a randomized experiment, so that we only observed data for households in treatment communities.

1. Do you think difference-in-means is a reasonable estimator of program impacts in this case? Why?

2. Is the Diff-in-diff estimator (with treated and untreated) any better? What assumptions underlie the use of this estimator?

3. Construct the difference-in-difference estimate of program impacts. How does it compare to that obtained using the experimental design?

(D) Non-experimental analysis: Propensity score matching. Suppose instead that we did not know the score nor the rule used by PROGRESA to allocate individuals to treatment and control status within treatment villages, we only observe recipients and non recipients.

1. Using available variables from the baseline, such as initial incomes, genders, and ages, construct an estimate of the propensity score using the stata command `pscore`. How does the choice of x variables affect calculation of the propensity score $p(x)$?

2. Inspect graphically the distribution of propensity scores for recipients and non recipients. Does it favor the overlap assumption?

3. Using this generated propensity score, estimate the ATT with Stata's commands `psmatch2`, using the default option (for nearest-neighbor matching) and the kernel option (for kernel matching). How do the estimates compare with each other? With the experimental results?

Chapter 2

Counterfactuals from the Real World: Difference-in-Differences and its Derivatives

Required Readings
Angrist and Pischke (2009): Chapter 5
Imbens and Wooldridge (2009): Section 6.5

Suggested Readings
Almond (2006)
Jensen (2007)
Beaman et al. (2012)
Heckman and Smith (1999)
Muralidharan and Prakash (2013)
Abadie et al. (2010)

2.1 Introduction

Sometimes we may be unwilling to assume that unconfoundedness holds, even after conditioning on covariates X. In this case we say there is *selection on unobservables*. This opens up an entirely new set of techniques which must be used to potentially estimate consistent effects of treatment. In section 2 and 3 we turn to these.

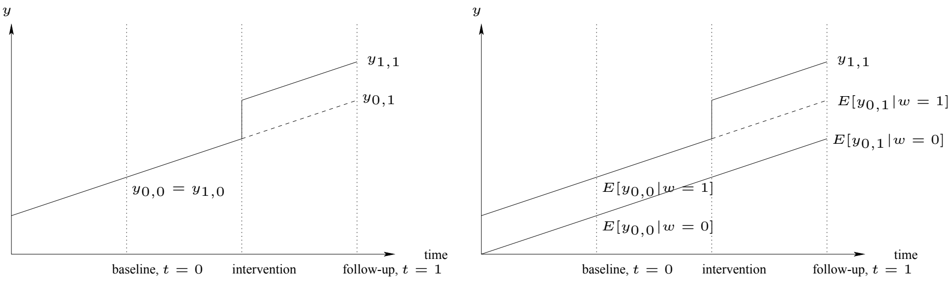
One particular case that we frequently encounter in which we may no longer believe in selection on unobservables is that of “natural experiments”, or naturally occurring events which strike certain units of a sample but not others. To take one simple example, imagine that we wish to estimate the impact of a natural disaster, such as an earthquake, on school completion rates. Earthquakes are to some degree geographically localized, and it seems reasonable to think that they are not endogenously determined by affected individuals. However, this does not suggest that it is appropriate to treat them as if they are randomly assigned, and simply compare the outcomes of affected individuals with those of unaffected individuals. One important factor is that an effect such as an earthquake hits an area which is potentially quite different at baseline to areas which are not hit by an earthquake, and as such, any difference following the earhtquake may owe to the event itself, or also to baseline differences in affected areas. In this case, the challenge in finding an appropriate counterfactual requires doing something to capture differences at baseline, and often methods such as “difference-in-differences” or related models are appropriate. We turn to these methods in this chapter.

2.2 Difference-in-Differences and Two-Way Fixed Effect Models

The basic underpinnings of the difference-in-differences (or diff-in-diff, or DiD, or DD, or double-difference) estimator is the case where we have observations of a pair of units across time, one of which is exposed to some policy or “treatment” of interest, and another of which is not. For example, for the case discussed above, imagine if we observe average highschool completion rates in two areas across two periods, and in one of the areas an earthquake occurred between the two periods, while in the other it did not. Even in the case that the standard assumption of unconfoundedness is not met (assumption 2 from the previous chapter), difference-in-differences allows us to recover an unbiased causal estimate if certain, weaker, assumptions are met.

Namely, difference-in-differences no longer requires that unconfoundedness holds, but does require that it holds in first differences, or that selection only owes to a fixed difference at baseline. Consider the following schematic example laid out in Figure 2.1. Here we use the notation $y_{treatment,time}$ to refer to the outcome depending on its exposure to the “treatment” (1 if eventually exposed, 0 if not), and time period (0 at baseline, 1 after treatment is in place). In the left-hand panel we can see what the requirements would be if wished to use a standard comparison of means type estimator based on unconfoundedness. Specifically, areas which are eventually treated and those which are eventually untreated must look identical. However, in the right hand panel we no longer have unconfoundedness, though it still holds in first differences given that there is a constant difference between eventually treated and eventually untreated units across time.[1] This right-hand case is suitable for diff-in-diff type methods, but would result in a clear bias if a comparison of means estimator was employed.

Figure 2.1: Panel Data: Levels and Difference



(a) Unconfoundedness holds

(b) Unconfoundedness fails in levels; holds in first differences

2.2.1 A Canonical Difference-in-Differences Set-up

Let's consider this two period and two area case and introduce some notation. We will refer to time periods t, and areas s to indicate states (though of course these may be regions, countries, villages, or even non-geographic units). Depending on the treatment status of an individuals state, We will thus observe one of two potential outcomes:

(a) y_{1ist} = Outcome for individual i at time t if their state of residence s is a treatment state

(b) y_{0ist} = Outcome for individual i at time t if their state of residence s is a non-treatment state.

As has always been the case with the potential outcomes, we will only observe at most one of these in a particular state and time period.

The diff-in-diff set-up assumes an additive structure for potential outcomes. We assume:

$$E[y_{0ist}|s,t] = \gamma_s + \lambda_t \tag{2.1}$$

This simply states that in the absence of treatment, the outcome consists of a time-invariant state effect (γ_s) and a year effect (λ_t) that is common across states.

We are interested in the effect of some treatment w, giving the potential outcome of:

$$y_{ist} = \gamma_s + \lambda_t + \tau w_{st} + \varepsilon_{ist}, \tag{2.2}$$

[1] This is the commonly referred to "parallel-trend assumption" in difference-in-differences which we will discuss further below.

where $E[\varepsilon_{ist}|s,t] = 0$. In what remains we will think of two states, which we'll call $Area^A$ and $Area^B$, and two time periods, which we'll call Pre and $Post$. In the Pre time period, neither state will receive treatment, however in the second time period treatement will "switch on" in $Area^A$.

Let's now consider what would happen if we were to estimate the treatment effect by comparing potential outcomes in both states in the $Post$ period:

$$\begin{aligned} E[y_{ist}|s = Area^A, t = Post] - E[y_{ist}|s = Area^B, t = Post] &= (\gamma^A + \lambda_{Post} + \tau) - (\gamma^B + \lambda_{Post}) \\ &= \tau + \gamma_A - \gamma_B. \end{aligned} \tag{2.3}$$

In this case, we would only recover the unbiased treatment effect in the particular case that the two states had identical mean values for γ, implying that they would have identical values of $E[y_{0ist}]$. Now, consider taking the first difference between the two states in the Pre period:

$$\begin{aligned} E[y_{ist}|s = Area^A, t = Pre] - E[y_{ist}|s = Area^B, t = Pre] &= (\gamma^A + \lambda_{Pre}) - (\gamma^B + \lambda_{Pre}) \\ &= \gamma_A - \gamma_B. \end{aligned} \tag{2.4}$$

Now, given that neither state receives treatment prior to the reform, all that remains is the baseline difference in $E[y_{0ist}]$. Then, combining these two single differences to form our double differences estimator gives:

$$\begin{aligned} E[y_{ist}|s = Area^A, t = Post] &- E[y_{ist}|s = Area^B, t = Post] - \\ E[y_{ist}|s = Area^A, t = Pre] &- E[y_{ist}|s = Area^B, t = Pre] = \\ (\tau + \gamma_A - \gamma_B) &- (\gamma_A - \gamma_B) = \tau. \end{aligned} \tag{2.5}$$

Thus, if our assumptions hold, diff-in-diff is a very elegant way to cancel out prevailing differences between treatment and control areas, and recover a causal estimate of treatment. These assumptions, of course, are something that we should always question. The key identifying assumption in the diff-in-diff world is the so called "parallel trends" assumption laid out in equation 2.1. In words, this just says that in the absence of treatment, all states would follow a similar trend, defined by γ_t. Treatment then induces a deviation from this common trend, as is illustrated in panel b of figure 2.1. These parallel trend assumptions are something that we spend a lot of time thinking about in diff-in-diff settings. We will return to this in section 2.2.4, and alternative specifications if we are not convinced in sections 2.2.6 and 2.3.

Estimating Difference-in-Differences

Fortunately, along with an elegant theoretical structure, this methodology is easy to take to data. Difference-in-differences can be very simply estimated in a regression framework. In order to do so, we generate a number of dummy variables to capture the additive structure defined in equation 2.2. Following the definitions above, we will define a dummy variable called "$Area^A$" which takes 1 if the individual lives in Area A, and 0 if they live in Area B.[2] Similarly, we will define a variable $Post$, which takes 1 during the second time period, and 0 in the first.

Now, to estimate our treatment effect of interest we simply perform the following regression:

$$y_{ist} = \alpha + \gamma Area^A_s + \lambda Post_t + \tau(Area^A_s \times Post_t) + \varepsilon_{ist}. \tag{2.6}$$

Our coefficient of interest τ, is associated with the term $Area^A \times Post$: the interaction term which switches on only in Area A after the reform. As Angrist and Pischke (2009, s. 5.2.1) lay out, this leads to the following interpretation of regression parameters:

$$\begin{aligned}
\alpha &= E[y_{ist}|s = Area^B, t = Pre] = \gamma^B + \lambda_{Pre} \\
\gamma &= E[y_{ist}|s = Area^A, t = Pre] - E[y_{ist}|s = Area^B, t = Pre] = \gamma^A - \gamma^B \\
\lambda &= E[y_{ist}|s = Area^B, t = Post] - E[y_{ist}|s = Area^B, t = Pre] = \lambda_{Post} - \lambda_{Pre} \\
\tau &= E[y_{ist}|s = Area^A, t = Post] - E[y_{ist}|s = Area^A, t = Pre] \\
&- E[y_{ist}|s = Area^B, t = Post] - E[y_{ist}|s = Area^B, t = Pre]
\end{aligned}$$

In this way, using a regression framework and appropriately defined dummy variables, we can immediately estimate both the desired treatment effect, as well as its standard error. This regression setup is extremely convenient for a few reasons:

1. The structure is very generalisable. In the examples so far, we have considered only a case where there are two states and two time periods. However, by including additional time dummy variables and additional state dummy variables in our regression model, we can extend this to a case with many states and/or many time periods. This is a frequently used estimation technique in the empirical economics literature. For example, the suggested reading of Almond (2006) provides a very nice example where many years of data are used, and many states are in both the treated and untreated groups. *However* a recently developing literature has shown significant challenges here when treatment effects are heterogeneous across groups or time. We return to this in section 2.2.2.

[2] Remember, given multicolinearity and the dummy variable trap, we only need 1 dummy variable if there are two geographical categories in the regression.

2. In this structure, we can replace our binary outcome "$Area^A$" for a variable indicating treatment intensity. For example, if treatment is not binary, with all states either being treated or un-treated, but rather varies by state, a measure of *intensity* can be used to replace $Area^A$ in the interaction term of 2.6. A classic example of this methodology is provided in Duflo (2001) (see her equation 1). We discuss this more formally when introducting "Fuzzy Diff-in-diff" models later in this chapter.

3. When we set up the conditional regression, there is nothing which stops us from controlling for additional (time varying) state-level variables. This allows us to control for things which we think may otherwise cause the parallel trends assumption not to hold. We will discuss this further in the next section.

Table 2.1: Regression Interpretation of Difference-in-Differences

Estimand	Estimate	
Panel A: Area A		
$E[y_{ist}\|s = Area^A, t = Post]$	$\alpha + \gamma + \lambda + \tau$	
$E[y_{ist}\|s = Area^A, t = Pre]$	$\alpha + \gamma$	
Single Difference = $(\alpha + \gamma + \lambda + \tau) - (\alpha + \gamma)$	=	$\lambda + \tau$
Panel B: Area B		
$E[y_{ist}\|s = Area^B, t = Post]$	$\alpha + \lambda$	
$E[y_{ist}\|s = Area^B, t = Pre]$	α	
Single Difference = $(\alpha + \lambda) - \alpha$	=	λ
Double Difference = $(\lambda + \tau) - \lambda$	=	τ

2.2.2 Two-Way Fixed Effect Models

Until quite recently, the two-by-two difference-in-differences model was treated as if it generally extended to multiple time periods and treated states, regardless of the context. For example, Angrist and Pischke (2009, p. 234) state "[i]t's also easy to add additional states or periods to the regression setup.", and Bertrand et al. (2004)'s seminal paper on inference in difference-in-differences models (which we will discuss later in this chapter) discuss the multi-period multi-state model as a "common generalization of the most basic DD setup". However, a growing body of work documents that these statements are only true if there is not heterogeneity in treatment effects estimated, for example if treatment effects are constant over time. This body of recent work, often with accompanying new methods and computational implementations, show that in the case of heterogeneous treatment effects, the standard "single coefficient" model may result in estimators which are quite different to what the model aims to capture. Here we discuss a number of these recent papers. In this whole section, we will consider a generalised model of the form of equation 2.2. Equation 2.2 assumes potentially multiple individual-level observations for each state s and time period t, and as such includes a subscript i. In many cases this is also estimated with a single average outcome for each state and year, in which case the specification can be simplified to:

$$y_{st} = \gamma_s + \lambda_t + \tau w_{st} + \varepsilon_{st}. \tag{2.7}$$

These two way fixed effect models are frequently encountered in empirical economics 'in the wild'. According to de Chaisemartin and D'Haultfoeuille (2020), 20% of the empirical papers published in the *American Economic Review* between 2010-2012 are based on this type of model.

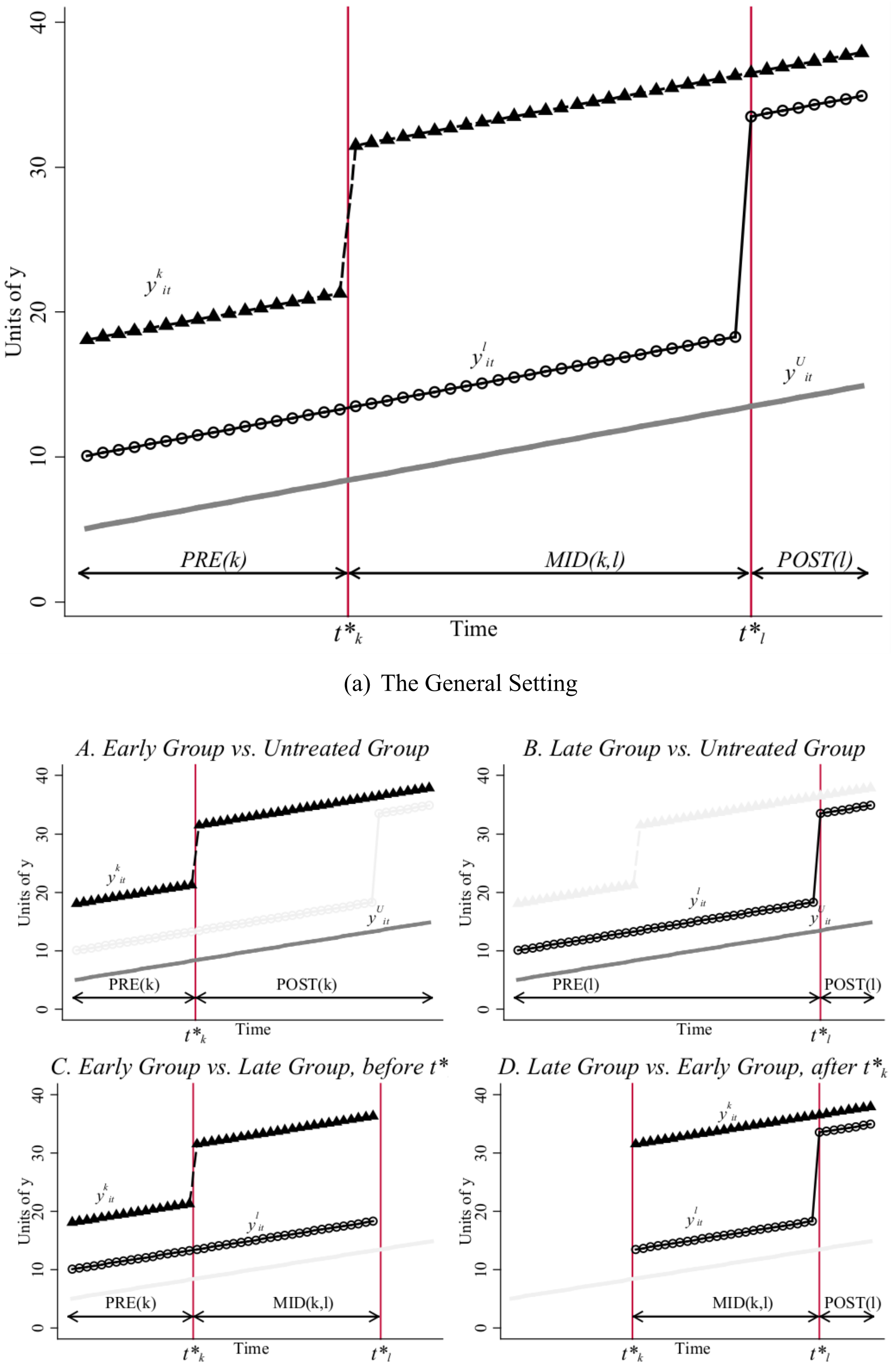
The Basic Idea

The basic concern with these models when there are mutliple time periods and multiple treatment states is that states may adopt treatment at different times. Athey and Imbens (2018) refer to these as "staggered adoption design". These staggered adoption designs can have important impacts on the nature of the coefficient estimated from equation 2.2 given that the key variation in estimating $\hat{\tau}$ comes from the moment when a unit changes treatment status, from non-treated to treated. Thus, if there are multiple periods, and a unit has already *changed* treatment status and is treated across multiple periods, given the nature of the OLS regression estimator it will itself be seen as a control unit in these periods given the lack of variation in w_{st} across these periods.

This has been laid out graphically in Goodman-Bacon (2021). The key graphs from this graphical set-up (Goodman-Bacon's Figures 1 and 2) are reproduced as Figure 2.2 below. In panel (a) we see an example where three states are considered (an early treatment indicated with triangles, a late treatment indicated with circles, and a never treated indicated as a solid line), with multiple time periods. As Goodman-Bacon (2021) (and others) show, the estimated parameter from 2.2 will consist of all possible combinations of "2×2" comparisons. These "2×2" comparisons are indicated in panel (b) using dark colours in each sub-panel. In particular, here there are two concerns. Firstly, there is one comparison which is somewhat strange, and that is the comparison indicated in D. Here the "control" group refers to the early treatment group, which does not change status in the post period, and the "treatment" group refers to the late treatment group which does change treatment status in this period. And the second issue is that these four "2×2" comparisons will not be given equal weight when arriving to a single coefficient estimate $\hat{\tau}$. We discuss these with more formal notation now.

Roughly speaking, the weighting issue is that each possible "2×2" comparison will receive a different weight, with this weight depending positively on the ... To see this formally, we

Figure 2.2: Goodman-Bacon (2021)'s Graphical Set-up with Three Treatment Groups



(a) The General Setting

(b) The DD Decomposition

lay out Goodman-Bacon (2021, proposition 1). We will use his notation for comparison with Figure 2.2. Consider a balanced panel of observations for $k = 1, \ldots, K$ groups receiving a binary treatment $t \in (1, T)$, as well as potentially an untreated group denoted U. The OLS estimate of equation 2.2 is a weighted average of all possible 2×2 DD estimators:

$$\widehat{\tau} = \sum_{k \neq U} W_{kU} \cdot \hat{\beta}^{2\times2}_{kU} + \sum_{k \neq U} \sum_{l > k} \left[W^k_{kl} \cdot \hat{\beta}^{2\times2,k}_{kl} + W^l_{kl} \cdot \hat{\beta}^{2\times2,l}_{kl} \right], \tag{2.8}$$

where here l refers to all units which adopt treatment later than unit k. Thus, the OLS estimate of the two-way fixed effect estimate actually consists of a weighted mean of three groups of estimates (the $\hat{\beta}$ terms refer to these estimates, and the weights are indicated by W). These estimates are precisely those laid out in panel (b) of figure 2.2. Specifically:

$$\hat{\beta}^{2\times2}_{kU} \equiv \left(\bar{y}^{Post(k)}_k - \bar{y}^{Pre(k)}_k\right) - \left(\bar{y}^{Post(k)}_U - \bar{y}^{Pre(k)}_U\right) \quad \text{[GROUPS A \& B]} \tag{2.9}$$

$$\hat{\beta}^{2\times2,k}_{kl} \equiv \left(\bar{y}^{Mid(k,l)}_k - \bar{y}^{Pre(k)}_k\right) - \left(\bar{y}^{Mid(k,l)}_l - \bar{y}^{Pre(k)}_l\right) \quad \text{[GROUP C]} \tag{2.10}$$

$$\hat{\beta}^{2\times2,l}_{kl} \equiv \left(\bar{y}^{Post(l)}_l - \bar{y}^{Mid(k,l)}_l\right) - \left(\bar{y}^{Post(l)}_k - \bar{y}^{Mid(k,l)}_k\right) \quad \text{[GROUP D].} \tag{2.11}$$

In understanding the global estimate, the question of interest relates to the weights given to each of the groups of estimates listed in equations 2.9, 2.10 and 2.11. These sum to one, and are:

$$W_{kU} = \frac{(n_k + n_U)^2 \hat{V}^D_{kU}}{\hat{V}^D} \tag{2.12}$$

$$W^k_{kl} = \frac{[(n_k + n_l)(1 - \bar{D}_l)]^2 \hat{V}^{D,k}_{kl}}{\hat{V}^D} \tag{2.13}$$

$$W^l_{kl} = \frac{[(n_k + n_l)\bar{D}_k]^2 \hat{V}^{D,l}_{kl}}{\hat{V}^D} \tag{2.14}$$

$$\tag{2.15}$$

where n refers to the sample share of a particular group in the whole, and $\bar{D}$ refers to the share of time that the sample is treated. Finally, the terms $\hat{V}^D_{kU}$, $\hat{V}^{D,k}_{kl}$ and $\hat{V}^{D,l}_{kl}$ refers to how much treatment varies in each subsample,[3] and $\hat{V}^D$ to how much treatment varies overall. This variance is largest when the two groups are closer in size and when treatment occurs closer to the middle of the considered time period.

While this weighting of the two-way fixed effect estimator is very interesting, what is perhaps more key is that where there is heterogeneity in the effects over time, the OLS estimate may be a considerably biased estimate of a weighted average of ATTs. While the full derivations are provided in Goodman-Bacon (2021, section II), the logic comes from the comparison between "already treated units" and newly treated units. If the impact of treatment is changing

[3] This depends positively on having groups which are more equally balanced between treatment and control units, and the variation in the treatment indicator in the subsample. See equations 7-9 of Goodman-Bacon (2021) for full definitions.

over time and already treated units are used as "control" units in future periods, this change in treatment effect will be mistakenly included as part of the control group. As Goodman-Bacon (2021) states, this "yields estimates that are too small or even wrong-signed".

A somewhat related discussion, along with a proposed alternative estimator, is provided in de Chaisemartin and D'Haultfoeuille (2020). We discuss their results, as well as their proposed "DID_M" estimator, in what follows. To do so, we (roughly)[4] follow their notation. That is, we consider a treatment applied at the level of group S and time T. For each $(s, t) \in \{1, \ldots, S\} \times \{1, \ldots, T\}$, the quantity $N_{s,t}$ refers to the number of observations in this group s, t, and the total quantity of observations is $N = \sum_{s=1}^{G} \sum_{t=1}^{t} N_{s,t}$. Note that in this case, if there is a single observation for each state and time period this is not an issue, but the design also allows for multiple observations in each group. For each $(i, s, t) \in \{1, \ldots, N_{s,t}\} \times \{1, \ldots, S\} \times \{1, \ldots, T\}$, the variable $D_{i,s,t}$ is the treatment status, and $(Y_{i,s,t}(0), Y_{i,s,t}(1))$ are potential outcomes without and with treatment respectively. Finally, $D_{s,t}$, $Y_{s,t}(0)$, $Y_{s,t}(1)$ and $Y_{s,t}$ all refer to simple averages over i.

de Chaisemartin and D'Haultfoeuille (2020) define $\widehat{\beta}_{fe}$ as the coefficient estimated in the following (standard) two-way fixed effects regression:

$$Y_{i,s,t} = \beta_0 + \beta_{fe} D_{s,t} + \mu_s + \lambda_t + \varepsilon_{s,t},$$

which is essentially what we define in equation 2.7. They also define the ATE for any (s,t) cell as:

$$\Delta_{s,t} = \frac{1}{N_{s,t}} \sum_{i=1}^{N_{s,t}} [Y_{i,s,t}(1) - Y_{i,s,t}(0)],$$

and define δ^{TR} as:

$$\delta^{TR} = E\left[\sum_{s,t:D_{s,t}=1} \frac{N_{s,t}}{N_1} \Delta_{s,t}\right], \tag{2.16}$$

where N_1 refers to the sum of all treated observations. This quantity δ^{TR} is the expectation of the weighted average of the ATE among all treated units, or in other words, the expectation of the ATT. One of the key results of de Chaisemartin and D'Haultfoeuille (2020) is to show that under a series of standard assumptions for difference-in-differences style models[5]:

$$\beta_{fe} = E\left[\sum_{s,t:D_{s,t}=1} \frac{N_{s,t}}{N_1} w_{s,t} \Delta_{s,t}\right], \tag{2.17}$$

[4] I have replaced g for groups with s for states, so that this notation follows more closely what we have been doing so far.

[5] These assumptions are (1) a balanced panel over s and t, (2) A "sharp" design where all units of a state receive treatment at the same time, (3) A no multi-colinearity assumption, (4) Strict exogeneity and (5) a parallel-trends assumption. If you wish to see how equality 2.16 is shown, refer to de Chaisemartin and D'Haultfoeuille (2020) proof of theorem 1 in their appendix A. This is not necessary for this course.

where:

$$w_{s,t} = \frac{\varepsilon_{s,t}}{\sum_{s,t:D_{s,t}=1} \frac{N_{s,t}}{N_1} \varepsilon_{s,t}}, \tag{2.18}$$

and $\varepsilon_{s,t}$ is the residual from a regression of $D_{s,t}$ on state and time fixed-effects. This is important, given that it implies that generally $\beta_{fe} \neq \delta^{TR}$, or in other words, $\hat{\beta}_{fe}$ is a biased estimator of the ATT. This is clear in comparing 2.16 with 2.17. The existence of the weighting term in 2.17 implies that certain groups' treatment effects will be given more or less weight. And what is most worrying with this result is that $w_{s,t}$ can be negative. In their Proposition 1 de Chaisemartin and D'Haultfoeuille (2020) show that these negative weights are more likely when:

- The ATE of interest is in a period in which a larger fraction of units are treated
- The ATE is for a group which is treated for many periods

A somewhat related decomposition is provided in Athey and Imbens (2018, Lemma 5) (who additionally go on to discuss a number of very important points on inference), however their baseline assumptions and final result are somewhat different. In section 3.1 of their paper de Chaisemartin and D'Haultfoeuille (2020) provide a simple numerical illustration. We consider a different example later on in this section of these notes.

What Should we do?

de Chaisemartin and D'Haultfoeuille (2020) propose an alternative estimator that is suitable for heterogeneous treatment effects. They define the quantity:

$$\delta^S = E\left[\frac{1}{N_S} \sum_{(i,s,t:t\geq 2, D_{s,t}\neq D_{s,t-1})} [Y_{i,s,t}(1) - Y_{i,s,t}(0)]\right]$$

where N_S refers to the quantity of treated units in the indicated cells. δ^S is thus the ATE for all cells that *change* their treatment status (eg from 0 to 1), at the moment that they begin to receive their new treatment. Under a series of assumptions laid out in their paper – including a common trends assumption[6] – thy suggest that this can be estimated using the following quantities. First define:

$$DID_{+,t} = \sum_{s:D_{s,t}=1, D_{s,t-1}=0} \frac{N_{s,t}}{N_{1,0,t}}(Y_{s,t} - Y_{s,t-1}) - \sum_{s:D_{s,t}=D_{s,t-1}=0} \frac{N_{s,t}}{N_{0,0,t}}(Y_{s,t} - Y_{s,t-1})$$

which compares the evolution of average outcomes in units changing treatment status between

[6] Specifically, they assume "Common trends for $Y(1)$", or that: for $t \geq 2$, $E(Y_{s,t}(1) - Y_{s,t-1})$ does not vary across s.

$t-1$ and t with those who remain unchanged. Similarly, if relevant,

$$DID_{-,t} = \sum_{s:D_{s,t}=D_{s,t-1}=1} \frac{N_{s,t}}{N_{1,1,t}}(Y_{s,t}-Y_{s,t-1}) - \sum_{s:D_{s,t}=0,D_{s,t-1}=1} \frac{N_{s,t}}{N_{0,1,t}}(Y_{s,t}-Y_{s,t-1})$$

captures the change between mean outcomes between $t-1$ and t comparing those units which stop receiving treatment to units whose treatment status remains unchanged. In the case of standard "staggered" two way FE models where units adopt treatment and are then treated forever after, $DID_{-,t}$ will not exist, and so will be set as 0 by definition. Then, they propose their DID_M estimator es as the weighted average of these quantities over all time periods:

$$DID_M = \sum_{t=2}^{T}\left(\frac{N_{1,0,t}}{N_S}DID_{+,t} + \frac{N_{0,1,t}}{N_S}DID_{-,t}\right).$$

They show that under their assumptions, $E[DID_M]=\delta^S$, a potentially more reasonable treatment estimator to consider. A similar estimator is proposed in Imai and Kim (2020). One benefit of this estimator is that we can use a similar version to consider both placebo tests, as well as dynamic treatment effects. de Chaisemartin and D'Haultfoeuille (2020) propose using the same structure of the DID_M estimator to estimate placebo versions, where rather than comparing changes between groups and $t-1$ and t, we compare lagged treatments, for example between $t-2$ and $t-1$. These are placebos as if we believe that parallel trends hold, we should see that these estimates *prior* to treatment do not result in any significant treatment effect. Similarly, we can consider dynamic treatment effects if rather than comparing changes between $t-1$ and t, we compare changes between the baseline difference, and other future periods, such as between $t-1$ and $t+1$, between $t-1$ and $t+2$, and so forth. They provide software to implement these estimators (and more), which we will discuss slightly later in this section.

Like de Chaisemartin and D'Haultfoeuille (2020), Callaway and Sant'Anna (2021) also start their analysis considering group and time specific treatment effects, focusing on $ATT(s,t)$ for groups (which here we denote s) at different time periods t. Based on these group-specific treatment effects, Callaway and Sant'Anna (2021) discuss "making inference about, …, funciontals of $ATT(s,t)$".[7] Among other things, starting with group and time specific treatment effects allows for the consideration of various types of heterogeneity, including heterogeneity by time of adoption, and by time since adoption.

While Callaway and Sant'Anna (2021) discuss a more complex estimator when controls are required, if a standard parallel-trends assumption holds, they note that

$$ATT(s,t) = E[Y_t - Y_{s-1}|G_s=1] - E[Y_t - Y_{s-1}|C=1].$$

[7] Callaway and Sant'Anna (2021) use the letter g to refer to groups, whereas here we are using s for consistency with earlier discussion.

where G_s is a binary indicator taking one if a state is first treated in period s. Thus, here states are indexed by their treatment time, called s. In the case that a state is never treated, it is included in controls, and the binary variable $C = 1$ for these units, and 0 otherwise. Thus, states must have one and only one value arcoss all variables G_s ($\forall s$) and C. They call this quantity $ATT(s,t)$ the "Group Average Treatment Effect" as it will (potentially) be different for each treatment group s and time period t. Their paper focuses on how to best aggregate these treatment effects in a logical way, and they suggest a range of estimators. Their most simple aggregates are:

$$\frac{2}{\mathcal{T}(\mathcal{T}-1)}\sum_{s=2}^{\mathcal{T}}\sum_{t=2}^{\mathcal{T}}1\{s \leq t\}ATT(s,t) \qquad \text{and} \qquad \frac{1}{\kappa}\sum_{s=2}^{\mathcal{T}}\sum_{t=2}^{\mathcal{T}}1\{s \leq t\}ATT(s,t)P(G=s)$$

where $\mathcal{T}$ is the final treatment period, and $1\{\cdot\}$ is a binary indicator implying the indicated condition should be met. Thus, these estimators thus weight all observed ATTs, either with the same weight for each group and year (left-hand case) or depending on their frequency in the data (right hand case). This avoids the weighting issue laid out by (among others) de Chaisemartin and D'Haultfoeuille (2020). However, Callaway and Sant'Anna (2021) also show that there are various other estimators that likely make sense, including estimates aggregated by:

- the period when units are first treated
- the amount of time the treatment has been in place (dynamic treatment effects)
- the effect in a given calendar year (calendar time effects).

In their paper they provide full derivations, estimation and inference details, and an illustration based on the effect of the US minimum wage and teen employment.

A Discussion on Practical Issues

Fortunately many of the recent advances in this literature come with accompanying computational routines, frequently written for R or for Stata. For example, Goodman-Bacon's decomposition has been implemented in Stata as bacondecomp (Goodman-Bacon et al., 2019), and a version is also available in R. The methods of de Chaisemartin and D'Haultfoeuille (2020) are available in Stata provided by de Chaisemartin et al. (2019b) (for implementing their proposed DID_M estimator) and de Chaisemartin et al. (2019a) (for calculating the weights of all treatment effects. The estimators proposed by Callaway and Sant'Anna (2021) are available in R as the "did" package (Callaway and Sant'Anna, 2020). These are a tremendous resource for simply implementing the latest methods designed for causal inference in models of these types, however it is important to understand the methods behind the libraries before simply diving into estimation. We will discuss a simple example further below.

A Numerical Example

The results from Goodman-Bacon and those of de Chaisemartin and D'Haultfoeuille (2020) are similar, however they take quite different paths to get there. Goodman-Bacon's (like that laid out in Athey and Imbens (2018)) is "mechanical" in that it is based on the underlying diff-in-diff comparisons. The result in de Chaisemartin and D'Haultfoeuille (2020) is based on a potential outcomes frame-work and a series of assumptions underlying the regression. This to examine how these methods work requires somewhat different frameworks. In the case of Goodman-Bacon (2021), we should consider all possible DD comparisons, while in the case of de Chaisemartin and D'Haultfoeuille (2020) we should consider each unit's ATE, which requires knowing the observed and counterfactual state. To examine this in a more applied way, let's consider a construted example.

Consider a panel of 3 states/areas, over the 10 years (t) of 2000 to 2009. One of these units is entirely untreated (unit=1), one is treated at time period 2003 (unit=2), and the other is treated at time period 2006 (unit=3). We will consider a simple-case where the actual data-generating process is known as:

$$y_{unit,t} = 2 + 0.2 \times (t - 2000) + 1 \times unit + \beta_1 \times post \times unit + \beta_2 \times post \times unit \times (t - treat).$$

In this mode $unit$ refers to the unit number listed above (1, 2 or 3), $post$ indicates that a unit is receiving treatment in the relevant time period t, and $treat$ refers to the treatment period (2003 for unit 2, and 2006 for unit 3). We will examine this set-up in R and Stata code in class.[8] This specification allows for each unit to have its own fixed effect, given that $unit$ is multiplied by 1, and allows for a general time trend increasing by 0.2 units each period across the whole sample. The impact of treatment comes from the units β_1 and β_2. The first of these, β_1, captures an immediate unit-specific jump when treatment is implemented which remains stable over time. The second of these, β_2, implies a trend break occurring *only* for the treated units once treatment comes into place. We will consider 2 cases below. In one case $\beta_1 = 1$ and $\beta_2 = 0$ (a simple case with a constant treatment effect per unit), and in a second case $\beta_1 = 1$ and $\beta_2 = 0.45$ (a more complex case in which there are heterogeneous treatment effects over time. These two cases are plotted in panels (a) and (b) respectively of Figure 2.3.

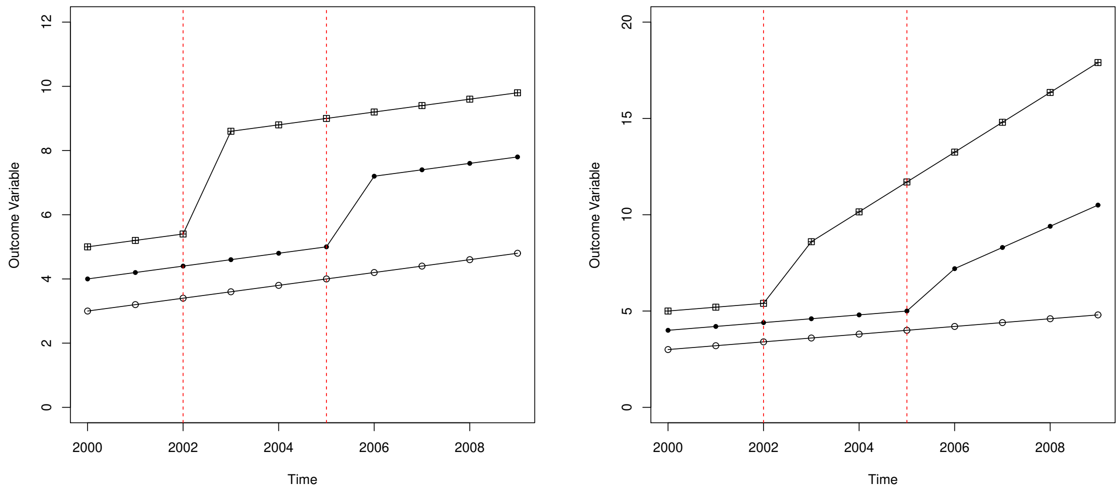
Below each panel of the plot we provide the decomposition of each treatment effect following the formulae of Goodman-Bacon (2021) and de Chaisemartin and D'Haultfoeuille (2020). Note that in the case of Goodman-Bacon (2021) this requires calculating four specific effects, which are the comparisons of each treated unit with the untreated unit, and each treated unit with each other. In the simple decomposition these are constant effects of 3 and 2 for early and later treated units given that the "treatment effect" is simply $1 \times unit$ in each case. However, in the second case this is more complicated, as we must take into account the time trends. This

[8] Refer to twowayfe.R and twowayfe.do.

results in the surprising behaviour flagged by Goodman-Bacon (2021) where despite each unit specific treatment effect being positive, the parameter $\widehat{\beta}_{kl}^{2\times 2,l}$ is actually *negative* given that it compares the change from the later-adopting unit (unit 2) with the unchanging portion of the earlier-adopting unit (unit 3), where the treatment effect for unit 3 grows *more* over time than that of unit 2.

Below the decomposition following Goodman-Bacon, we present the decomposition of de Chaisemartin and D'Haultfoeuille (2020). Here we must calculate an ATE for each unit which recevies treatment in each period, where the counterfactual case simply refers to the outcome had β_1 and β_2 been 0. Here, we once again see why the total treatment effect (called β_{fe}) in de Chaisemartin and D'Haultfoeuille (2020) is not a simple average of all unit-specific treatment effects. Given the proportions of treated and untreated observations for each unit, the post-2006 ATEs for unit 3 are given 0 weights, and hence form no part of the global ATT. Note that if we change the time period when units first receive treatments, this weight can even turn negative (for example if unit 3 first receives treatment prior to 2003, or unit 2 first receives treatment after 2006). We will explore this in more depth in the R and Stata codes, which also examine how the estimators proposed by de Chaisemartin and D'Haultfoeuille (2020) result in a much more logical treatment effect.

Figure 2.3: A Numerical Example of Time-Varying Treatment Effects



(a) Simple Decomposition (b) Decomposition with trends

	(a) Simple Decomposition		(b) Decomposition with trends	
	$\hat{\beta}$	Weight	$\hat{\beta}$	Weight
Goodman-Bacon				
$\widehat{\beta}_{kU}^{2\times 2}$	3	0.318	7.05	0.318
$\widehat{\beta}_{lU}^{2\times 2}$	2	0.364	3.35	0.364
$\widehat{\beta}_{kl}^{2\times 2,k}$	3	0.136	4.35	0.136
$\widehat{\beta}_{kl}^{2\times 2,l}$	2	0.182	-1.38	0.182
$\widehat{\tau}$	2.45	1	3.8	1
de Chaisemartin and D'Haultfoeuille				
$\widehat{\beta}_{2,2006}$	2	0.136	2	0.136
$\widehat{\beta}_{2,2007}$	2	0.136	2.9	0.136
$\widehat{\beta}_{2,2008}$	2	0.136	3.8	0.136
$\widehat{\beta}_{2,2009}$	2	0.136	4.7	0.136
$\widehat{\beta}_{3,2003}$	3	0.152	3	0.152
$\widehat{\beta}_{3,2004}$	3	0.152	4.35	0.152
$\widehat{\beta}_{3,2005}$	3	0.152	5.7	0.152
$\widehat{\beta}_{3,2006}$	3	0	7.05	0
$\widehat{\beta}_{3,2007}$	3	0	8.4	0
$\widehat{\beta}_{3,2008}$	3	0	9.75	0
$\widehat{\beta}_{3,2009}$	3	0	11.1	0
$\widehat{\beta}_{fe}$	2.45	1	3.8	1

2.2.3 Inference in Diff-in-Diff

Up to this point, we have principally focused on estimation of difference-in-differences and two-way fixed effect models. However, there is a long literature pointing to important inferential considerations which must be taken into account for the estimation of appropriate standard errors and the construction of appropriate confidence intervals.

A Brief Review of Variance and Standard Error Estimation As you will remember from prior econometrics courses, estimating standard errors correctly relies on the Gauss-Markov assumptions. However, in many cases, it is hard to assume that the ε_{it} terms are i.i.d. For one, we may expect that individual outcomes in the same area and the same year may be correlated: $Cov(\varepsilon_{it}, \varepsilon_{jt} | s_i = s_j) \neq 0$. We would also expect shocks which affect each group to be serially correlated over time ($Cov(\varepsilon_{it+1}, \varepsilon_{jt} | s_i = s_j) \neq 0$). Bertrand et al. (2004) discuss many of these issues in their paper "How Much Should We Trust Differences-in-Differences Estimates?".

The most commonly used solution to this problem is to cluster standard errors at the group level. To see how this works, let's start with the most basic "plain vanilla" standard errors. As you will likely recall, we calculate standard errors from the variance-covariance matrix of our OLS estimators β. In particular, the standard errors are the square root of the variance of a particular estimator (or the square root of the diagonal of the variance-covariance matrix). For now, let's consider a simple model with a single independent variable x_i and a dependent variable y_i, each of which have been demeaned for simplicity. We can thus write the formula for the variance-covariance matrix as follows:

$$V(\hat{\beta}) = \frac{V\left[\sum_{i=1}^{N} x_i u_i\right]}{\left(\sum_{i=1}^{N} x_i^2\right)^2}$$

where u_i refers to the residual term of our OLS model.

Of course we never actually observe u_i, so to arrive at an estimable variance-covariance matrix we need to go slightly further. In the simplest case where we assume that errors are completely uncorrelated, the numerator of this variance-covariance matrix is: $V\left[\sum_{i=1}^{N} x_i u_i\right] = \sum_{i=1}^{N} V[x_i u_i] = \sum_{i=1}^{N} x_i^2 V[u_i] = \sum_{i=1}^{N} x_i^2 \sigma^2$, and the variance is thus:

$$\widehat{V}(\hat{\beta}) = \frac{\widehat{\sigma}^2}{\sum_{i=1}^{2} x_i^2}.$$

Note in the above that now we add a hat to the V term as it is an estimated quantity, and that this estimate depends on σ^2, which is estimated by OLS as $\widehat{\sigma}^2 = \sum_{i=1}^{N} \hat{u}_i^2/(N-K)$.

This most basic calculation for $\widehat{V}(\hat{\beta})$ assumes that the variance of u_i is constant for all observations (homoscedasticity). From introductory econometrics courses we already know of one type of loosening of this most basic variance-covariance matrix, and this is the heteroscedasticity robust version of White (1980).

$$\widehat{V}(\hat{\beta})_H = \frac{\left(\sum_{i=1}^N x_i^2 \hat{u}_i^2\right)}{\left(\sum_{i=1}^2 x_i^2\right)^2}.$$

In the above we add a subscript H to indicate that it is heteroscedasticity robust, where we note that we now allow arbitrary correlations between x_i and u_i in the numerator term.

However, what we want with clustered standard errors is not that an individual's error term can depend on their own level of x_i, but rather that the error of one individual can be correlated with error of another individual! So then, we need to allow for a further loosening of the variance-covariance matrix to build in this cross-unit dependence. This brings us to the cluster-robust version:

$$\widehat{V}(\hat{\beta})_C = \frac{\left(\sum_{i=1}^N \sum_{j=1}^N x_i x_j \hat{u}_i \hat{u}_j \mathbb{1}\{i,j \text{ from the same cluster}\}\right)}{\left(\sum_{i=1}^2 x_i^2\right)^2}.$$

In this formula, $\mathbb{1}$ is an indicator function equal to one if two individuals share a cluster, and 0 otherwise. This avove variance-covariance matrix now permits not only homoscedasticity, but also arbitrary correlation between units *within* clusters. This is what we generally prefer to use in difference-in-difference estimates.

This variance term is typically larger than the "standard" variance term of OLS, and the degree to which the cluster robust variance inflates the (overly precise) standard variance is known as the "Moulton factor", after the paper which laid this out (Moulton, 1986). For additional discussion, see Moulton (1986) or Angrist and Pischke (2009, chapter 8), but note that in general, the downward bias in the standard OLS variance general increases to the degree that:

(a) The size of clusters is larger

(b) The correlation within the cluster for the variable of interest

(c) The correlation within the cluster of the error term.

Practical Considerations in Variance Estimates This formula is presented in matrix form in Cameron et al. (2008). Consider a regression model where observations i are clustered in groups s:

$$y_{is} = \boldsymbol{x}'_{is}\boldsymbol{\beta} + u_{is} \quad \text{for} \quad i = 1, \ldots, N, \quad s = 1, \ldots, S.$$

This can be aggregated to the level of the cluster as:

$$\boldsymbol{y}_s = \boldsymbol{X}_s\boldsymbol{\beta} + \boldsymbol{u}_s, \quad s = 1, \ldots, S,$$

or simply in matrix form as:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u},$$

Here, the cluster-robust variance-covariance estimator (CRVE) can be written in matrix form as:

$$\widehat{V}_{CR}(\widehat{\beta}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\left(\sum_{s=1}^{S}\boldsymbol{X}_s\hat{\boldsymbol{u}}_s\hat{\boldsymbol{u}}_s'\boldsymbol{X}_s'\right)(\boldsymbol{X}'\boldsymbol{X})^{-1}, \tag{2.19}$$

where $\hat{\boldsymbol{u}}_s\hat{\boldsymbol{u}}_s'$ estimates the intra-cluster correlation. The standard solution for difference-in-difference style models is to allow for within-cluster auto-correlation by using a CRVE such as the above to estimate standard errors and confidence intervals on regression parameters. Such an estimator is provided as standard in Stata by specifying the `vce(cluster` *clustvar*`)` option in regression models. However, as has been extensively documented (eg Cameron and Miller (2015)), standard CRVEs are only asymptotically valid, where the asymptotic behavior depends on the number of clusters $S \to \infty$. When standard clustering is used based on 'too few' clusters, the CRVE is generally downward-biased, resulting in over-rejection of null hypotheses. This bias occurs because $E(\hat{\boldsymbol{u}}_s\hat{\boldsymbol{u}}_s') \neq \boldsymbol{u}_s\boldsymbol{u}_s'$ in equatiuon 2.19, with the latter term being the true intra-cluster variation. While in general, computational packages make small sample corrections for this bias[9], in certain cases this bias can be severe (Cameron and Miller, 2015; Mackinnon and Webb, 2018), even using these standard corrections.

Thus, while clustering is computationally simple, clustered standard errors are generally only correct if "enough" clusters are included. This implies that for clustered standard errors to hold in diff-in-diff regressions, a sufficient number of treatment and non-treatment states must exist. In practice, knowing how many clusters it 'too few' depends on a number of factors. While there are rules of thumb such as the rule of 42 laid out in Angrist and Pischke (2009) which suggests that standard clustering provides a good approximation if $S \geq 42$ clusters, the performance of these methods under simulation has been shown to depend also on the relative size of clusters (Mackinnon and Webb, 2017). A range of results surveyed in Cameron and Miller (2015) leads to their suggestion that if one is analyzing data with fewer than 50 clusters in a state-year panel (such as the case with panel-event studies), alternative inference methods should be considered.

[9] For example, Stata estimates the CRVE as:

$$\widehat{V}_{CR}(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\left(\sum_{s=1}^{S}\boldsymbol{X}_s\tilde{\boldsymbol{u}}_s\tilde{\boldsymbol{u}}_s'\boldsymbol{X}_s'\right)(\boldsymbol{X}'\boldsymbol{X})^{-1},$$

where $\tilde{\boldsymbol{u}}_s = \sqrt{c}\hat{\boldsymbol{u}}_s$, with c being a small sample correction $c = (S/(S-1)) \times ((N-1)/(N-k))$ and $\hat{\boldsymbol{u}} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ are standard regression residuals (Cameron et al., 2008).

However, what should we do if we have an application in which we would like to cluster our standard errors, but don't have a large enough number of clusters? It is important to note that the answer most certainly *is not* 'just cluster anyway'. If we use traditional clustered standard errors with a small number of clusters we will very likely underestimate our standard errors, and thus over-reject null hypotheses. Fortunately, alternative solutions do exist. The most common solution is to use a wild cluster bootstrap. This is based on the logic of the bootstrap. The bootstrap, from Efron (1979) is a resampling procedure. In this case, rather than calculating standard errors analytically (ie using a formula), we simulate many different samples of data, and based on estimates from each sample we can observe the variation in underlying parameters of interest, and hence build confidence intervals and rejection regions. The idea of the bootstrap is that we should treat the sample as the population. Then we can draw (with replacement) many samples of size N from this "population", and for each of these resamples we can calculate our estimator of interest, arriving at a distribution for the estimator and hence confidence intervals and standard errors. The wild bootstrap is simply a type of bootstrap procedure where we resample respecting the original clusters in our data. We will discuss this at more length in an example in class.

In this case where the quasi-experimental set-up is based on fewer than around 50 clusters, the wild cluster bootstrap has been documented to be a successful resampling-based method to take account of auto-correlation in variables underlying panel event studies, even in cases with fewer clusters (see *eg* Cameron et al. (2008); Cameron and Miller (2015); Roodman et al. (2019)). This has been efficiently implemented in Stata as described in Roodman et al. (2019), and programmed for Stata as `boottest` (Roodman, 2015). Finally, note that in the case of very few clusters, and in particular few clusters where an event occurs, inference is further complicated. In cases such as this a number of potential solutions have been proposed, such as those described in Mackinnon and Webb (2018); Conley and Taber (2011). If you are interested in further details these papers will provide a comprehensive background.

2.2.4 Testing Diff-in-Diff Assumptions

In the preceding sections, we have seen that inferring causality in two-way fixed effect and difference-in-difference models relies crucially on the validity of the parallel trends assumption. If average outcomes in treatment and control areas would have followed different trends even in the absence of treatment, any estimated parameters will reflect both variations in prevailing trends, as well as the true treatment effect. While this is not something that we can ever test formally given that it requires observing the (unobserved) counterfactual state, one thing that we can sometimes do is formally examine how trends in outcomes treated and untreated areas were evolving *prior* to the reform. While this does not amount to a formal test of the parallel trend assumption, it would cast suspicion on our model assumptions if parallel trends did not even hold in the pre-treatment window.

One particular case in which these parallel pre-trends will not hold is the case of the so-called "Ashenfelter dip". This Ashenfelter dip, named for the labour economist Orley Ashenfelter, and particularly the results in Ashenfelter (1978), recognises that often participants in labour market training programs have a reduction earnings immediately *before* participation in the program. The logic of this is that if individuals self-select into training programs, many of those who select in will be those who have lost their job, and hence particiapte in the training program as part of a job search. This pattern of outcomes has been shown in a wide array of labour market training programs (see, for example, (Heckman and Smith, 1999)). The trouble with this sort of dynamics is that these reductions in mean salary are largely transitory, and the participants would have experienced an increase in salary in the following years even in the absence of the program. In other words, participants and non-participants *would not* have followed parallel trends, as participants should recover their earlier earnings, while non-participants face no such dynamic.

There exist a number of ways to examine the validity of the parallel trends assumption, which will identify, among other things, the Ashenfelter dip. However, even using these techniques, in no case can we ever *prove* definitively that it holds; we can only provide evidence suggestive that it is a reasonable assumption to make[10]. You could think of tests of this type as analogous to tests of instrumental overidentification. While they are not definitive proofs of assumptions, they at least provide some evidence that they aren't entirely unreasonable in the context examined.

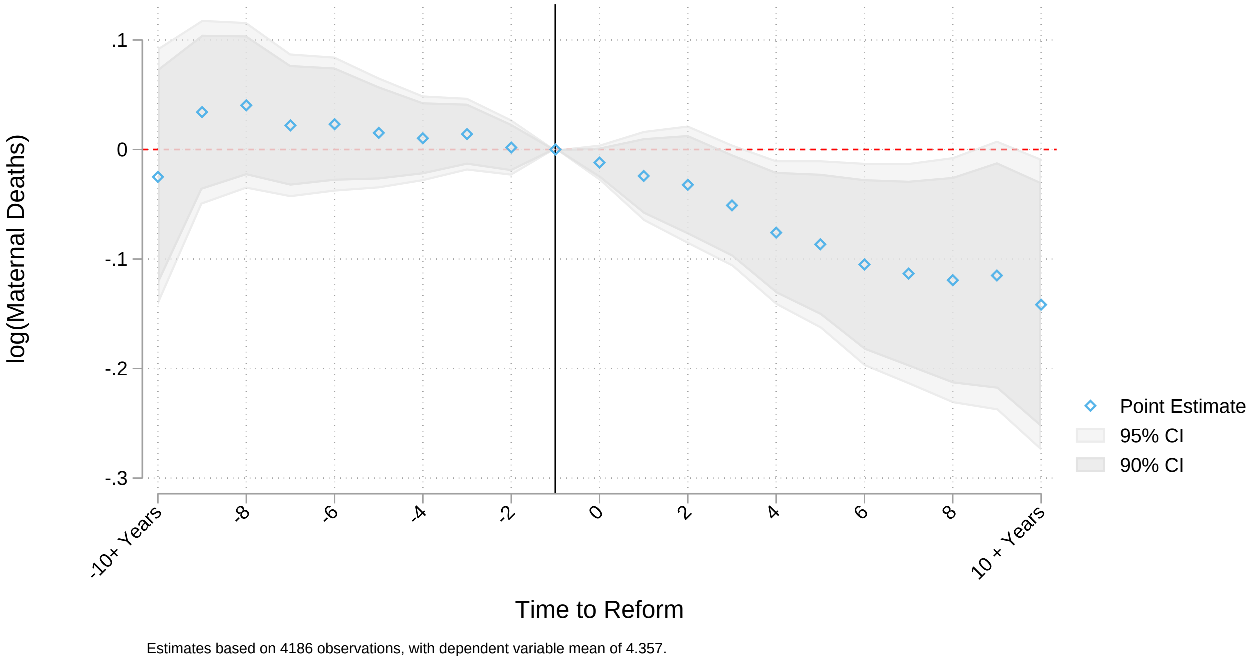
If multiple pre-treatment periods of data are observed, the simplest test is to remove all post-treatment data, and run the same specification, but using a placebo which tests whether any differences are found between treatment and control states entirely *before* the reform had been implemented. If we do find that there is a difference over time even in the absence of the reform, this may be quite concerning when moving to the post-reform case. A more extensive test of the validity of the parallel trends assumption is the use of the "panel event study analysis", which additionally allows us to examine the dynamics of any treatment effects post-reform.

An event study can be thought of as a test following the ideas of Granger Causality (Granger, 1969). If it is the case that the reform is truly causing the effect, we should see that any differences between treatment and control states emerge *only* after the reform has been implemented, and that in all years prior to the reform, differences between treatment and non-treatment areas remain constant. Thus, the basic idea behind the panel event study is that we should observe how outcomes in treated versus untreated areas evolve in the pre-reform period (relative to a baseline omitted category), as well as how the evolve once the reform has been put in place.

[10] This is just another example of the "fundamental problem of causal inference" of Holland (1986) that we discussed earlier in this lecture series.

A graphical example of what these sorts of test look like is provided below in figure 2.4. This presents rates of maternal mortality surrounding the adoption of a parliamentary gender quota (see Bhalotra et al. (2020)). In this case in each pre-reform period, no difference is observed between trends of maternal mortality in treatment and non-treatment areas. Following the reform however, a significant reduction in the outcome variable is seen in the treatment areas when compared to non-treatment areas. Results of this type provide significant support for the validity a difference-in-difference methodology, with the added benefit that we can also consider the dynamics of the effect of the reform over time. We turn to the specifics of this design in the following section.

Figure 2.4: Event Study Graph and Reform Timing



2.2.5 The Panel Event Study Model

There is a burgeoning literature discussiong panel event study methods, including the work of Borusyak and Jaravel (2018); Freyaldenhoven et al. (2019); Schmidheiny and Siegloch (2019). The discussion in this section is drawn from Clarke and Tapia Schythe (2020), which provides background, and a review of estimation in Stata. In laying out the panel event study, consider a panel covering states s and time periods t. We are interested in estimating the impact of the passage of an event which may occur at different times in different states (what we have been calling a staggered assignment design above). We will denote as $Event_s$ a variable recording the time period t in which the event is adopted in state s. Denoting the outcome of interest as

y_{st}, the panel event study specification can be written as[11]:

$$y_{st} = \alpha + \sum_{j=2}^{J} \beta_j (\text{Lag } j)_{st} + \sum_{k=1}^{K} \gamma_k (\text{Lead } k)_{st} + \mu_s + \lambda_t + X'_{st}\Gamma + \varepsilon_{st}. \tag{2.20}$$

Here μ_s and λ_t are state and time fixed effects, X_{st} are (optionally) time-varying controls, and ε_{st} is an unobserved error term. In equation 2.20, lags and leads to the event of interest are defined as follows:

$$\begin{aligned}
(\text{Lag } J)_{st} &= \mathbb{1}[t \leq Event_s - J], & (2.21)\\
(\text{Lag } j)_{st} &= \mathbb{1}[t = Event_s - j] \text{ for } j \in \{1, \ldots, J-1\}, & (2.22)\\
(\text{Lead } k)_{st} &= \mathbb{1}[t = Event_s + k] \text{ for } k \in \{1, \ldots, K-1\}, & (2.23)\\
(\text{Lead } K)_{st} &= \mathbb{1}[t \geq Event_s + K]. & (2.24)
\end{aligned}$$

Lags and leads are thus binary variables indicating that the given state was a given number of periods away from the event of interest in the respective time period. J and K lags and leads are included respectively, and, as indicated in equations 2.21 and 2.24, final lags and leads "accumulate" lags or leads beyond J and K periods. A single lag or lead variable is omitted to capture the baseline difference between areas where the event does and does not occur. In specification 2.20, as standard, this baseline omitted case is the first lag, where $j = 1$.

A stylized example of such a setting is provided in Table 2.2. We consider four states forming a balanced panel of years from 2000-2009. The $Event_s$ variable occurs at different times in different states, and in the case of one state, does not occur. Here both four lags and four leads are included, such that $J = K = 4$. Lag and Lead 4 (exclusively) are switched on for periods in which the "Time to Event" exceeds 4 lags or leads respectively.

[11] There are a number of ways to specify such a model. Slightly different notations are used by Schmidheiny and Siegloch (2019) who define the model as:

$$y_{st} = \sum_{j=\underline{j}}^{\overline{j}} \beta_j b_{st}^j + \mu_s + \lambda_t + \varepsilon_{st},$$

where

$$b_{st}^j = \begin{cases} \mathbb{1}[t \leq Event_s + j] & \text{if } j = \underline{j} \\ \mathbb{1}[t = Event_s + j] & \text{if } \underline{j} < j < \overline{j} \\ \mathbb{1}[t \geq Event_s + j] & \text{if } j = \overline{j}, \end{cases}$$

and where $\underline{j}$ is equivalent to our definition of J and $\overline{j}$ is equivalent to our L. In the case of Freyaldenhoven et al. (2019), they define a version of this model as:

$$y_{st} = \delta_{-K+}(1 - z_{s,t+(K-1)}) + \delta_{L+}z_{s,t-L} + \sum_{k=-(L-1)}^{K-1} \delta_{-k}\Delta z_{s,t+k} + \mu_s + \lambda_t + \varepsilon_{st},$$

where $z_{st} \equiv PostEvent_{st}$ as defined in Table 2.2, $z_{s,t+k}$ and $z_{s,t-k}$ refer to lags and leads of this variable respectively, and Δ refer to the first difference of these lag/lead terms. These models, and that laid out in equations 2.20-2.24 are equivalent.

Table 2.2: A Stylized Example of a Panel Event Study

State	Year	Event	Post	Time to	Lag	Lag	Lag	Lag	Lead	Lead	Lead	Lead	Lead
(s)	(t)		Event	Event	4	3	2	1	0	1	2	3	4
State A	2000	2004	0	-4	1	0	0	0	0	0	0	0	0
State A	2001	2004	0	-3	0	1	0	0	0	0	0	0	0
State A	2002	2004	0	-2	0	0	1	0	0	0	0	0	0
State A	2003	2004	0	-1	0	0	0	1	0	0	0	0	0
State A	2004	2004	1	0	0	0	0	0	1	0	0	0	0
State A	2005	2004	1	1	0	0	0	0	0	1	0	0	0
State A	2006	2004	1	2	0	0	0	0	0	0	1	0	0
State A	2007	2004	1	3	0	0	0	0	0	0	0	1	0
State A	2008	2004	1	4	0	0	0	0	0	0	0	0	1
State A	2009	2004	1	5	0	0	0	0	0	0	0	0	1
State B	2000	2005	0	-5	1	0	0	0	0	0	0	0	0
State B	2001	2005	0	-4	1	0	0	0	0	0	0	0	0
State B	2002	2005	0	-3	0	1	0	0	0	0	0	0	0
State B	2003	2005	0	-2	0	0	1	0	0	0	0	0	0
State B	2004	2005	0	-1	0	0	0	1	0	0	0	0	0
State B	2005	2005	1	0	0	0	0	0	1	0	0	0	0
State B	2006	2005	1	1	0	0	0	0	0	1	0	0	0
State B	2007	2005	1	2	0	0	0	0	0	0	1	0	0
State B	2008	2005	1	3	0	0	0	0	0	0	0	1	0
State B	2009	2005	1	4	0	0	0	0	0	0	0	0	1
State C	2000	.	0	.	0	0	0	0	0	0	0	0	0
State C	2001	.	0	.	0	0	0	0	0	0	0	0	0
State C	2002	.	0	.	0	0	0	0	0	0	0	0	0
State C	2003	.	0	.	0	0	0	0	0	0	0	0	0
State C	2004	.	0	.	0	0	0	0	0	0	0	0	0
State C	2005	.	0	.	0	0	0	0	0	0	0	0	0
State C	2006	.	0	.	0	0	0	0	0	0	0	0	0
State C	2007	.	0	.	0	0	0	0	0	0	0	0	0
State C	2008	.	0	.	0	0	0	0	0	0	0	0	0
State C	2009	.	0	.	0	0	0	0	0	0	0	0	0
State D	2000	2007	0	-7	1	0	0	0	0	0	0	0	0
State D	2001	2007	0	-6	1	0	0	0	0	0	0	0	0
State D	2002	2007	0	-5	1	0	0	0	0	0	0	0	0
State D	2003	2007	0	-4	1	0	0	0	0	0	0	0	0
State D	2004	2007	0	-3	0	1	0	0	0	0	0	0	0
State D	2005	2007	0	-2	0	0	1	0	0	0	0	0	0
State D	2006	2007	0	-1	0	0	0	1	0	0	0	0	0
State D	2007	2007	1	0	0	0	0	0	1	0	0	0	0
State D	2008	2007	1	1	0	0	0	0	0	1	0	0	0
State D	2009	2007	1	2	0	0	0	0	0	0	1	0	0

Example provided in Clarke and Tapia Schythe (2020).

States in which the event never occurs (such as State C in Table 2.2) act as pure controls. These units have 0s in all lag and lead terms, and act as the counterfactual on which the estimation of impacts is based. Differences between these pure controls states and states which adopt the event of interest are anchored at 0 in the omitted base period, ie the first lag in equation 2.20. Hence, lags and leads capture the difference between treated and control states, compared to the prevailing difference in the omitted base period. Unbiased estimation of post-event treatment effects thus relies fundamentally on the so called "parallel trends assumption". In the absence of treatment, it is assumed that treated and control states would have maintained similar differences as in the baseline period. For this reason, these models have been demonstrated to be under-identified, or identified only up to a linear trend, when all units adopt treatment at some point in time (Schmidheiny and Siegloch, 2019; Borusyak and Jaravel, 2018). Schmidheiny and Siegloch (2019) show that in this case, it is necessary to bin lags and leads beyond certain maximum lag (J) and lead (K) periods.

The panel event study is an extension of the standard two-way fixed effect (sometimes called difference-in-differences) model, where a single "Post Event" indicator is included for all periods posterior to the occurrence of the event in treated states. This is simply:

$$y_{st} = \alpha + \beta \text{PostEvent}_{st} + \mu_s + \lambda_t + X'_{st}\Gamma + \varepsilon_{st}, \tag{2.25}$$

where following the notation from (2.21)-(2.24), $\text{PostEvent}_{st} = \mathbb{1}[t \geq Event_s]$. Estimation of event specification 2.20 provides two key pieces of information not observable in this single-coefficient model. Firstly, the full set of event leads allows for the inspection of parallel trends in the *pre*-treatment period. While this does not provide evidence that the units in which the event was adopted and not adopted would have necessarily followed similar trends in the post-reform period (Kahn-Lang and Lang, 2019) (which is the identifying assumption of these models), if trends in treated and untreated areas were not parallel even pre-event, it is unlikely that they would be parallel post-event. Secondly, the policy lags allow for inspection of the temporal nature of treatment effects, noting for example, any dynamics in the appearance of effects, for example growing or shrinking over time, and whether effects are transitory or permanent.

While the results from papers such as Goodman-Bacon (2021) suggest that the esimation of panel event studies resolves concerns owing to heterogeneity in treatment effects and staggered adoption designs, results from Sun and Abraham (2020) suggest that specific types of heterogeneity concerns remain even in panel event study models. In particular, they note undesired weighting of treatment effects if there is heterogeneity across treatment groups in particular lag and lead terms. Other concerns exist in event study designs, such as possible inferential problems related to selective survival of models based on pre-trend tests (Roth, 2019). It is worth noting that along with the estimators discussed earlier in this section of de Chaisemartin and D'Haultfoeuille (2020); Callaway and Sant'Anna (2021), there are other alternatives including the stacked DD procedure of Sun and Abraham (2020), and sensitivity tests related to these panel event studies described in Roth (2019); Rambachan and Roth (2019) (with accompanying R code).

2.2.6 Other Extensions to Diff-in-Diff Methods

Difference-in-Difference-in-Differences

Difference-in-differences estimates frequently provide a good test for the impact of some reform. However, what can we do if we think that simply capturing a base-line difference in treatment and non-treatment areas is not enough? One option is to extend a the diff-in-diff approach to a diff-in-diff-in-diff (triple differences) approach! This follows the logic of difference-in-differences, however estimates the diff-in-diff model for two groups: one which is affected by the reform and one which isn't. If the group which is not affected by the reform has any change over time, this is then substracted from the main diff-in-diff estimate to give a triple difference estimate.

Perhaps the best way to think of this is to examine an applied example. Muralidharan and Prakash (2013) estimate a triple differences framework to estimate the effect of a program in the state of Bihar, India which gave girls (but not boys) funds to buy a bike to travel to school. As they point out, the logical difference in difference approach is to compare the change in enrollment rates of girls in Bihar before and after treatment with the change in enrollment rates of boys in the same state. This precisely follows the logic of the previous section of two groups in two locations. However, they are concerned that boys and girls were following different trends in highschool enrollment rates even *before* the reform. In order to control for this, they thus estimate the same regression in two states: Bihar, the treatment state, and Jharkhand, a nearby but untreated state.

The mechanics of actually estimating a regression are similar to specification 2.6, however now must account for the *triple* interactions. Defining subscript g to refer to gender now, they thus estimate:

$$\begin{aligned} y_{isgt} &= \beta_0 + \beta_1 Bihar_s + \beta_2 Girl_g + \beta_3 Post_t + \beta_5 (Bihar_s \times Girl_g) + \\ &\quad \beta_6 (Bihar_s \times Post_t) + \beta_7 (Post_t \times Girl_g) + \tau (Bihar_s \times Girl_g \times Post_t) + \varepsilon_{ist}. \end{aligned} \tag{2.26}$$

In this case, the treatment effect τ is captured by the triple interaction term. If you find all these interactions hard to follow, you may want to figure out what each coefficient is capturing as per the system of coefficient equations laid out in section 2.2.1!

"Fuzzy" Differences-in-Differences

So far, when discussing difference-in-differences and two-way fixed effect methods, we have assumed that all units of a state receive treatment at a given point in time, and other states exist which never receive treatment to act as control states. However, as (de Chaisemartin and D'Haultfoeuille, 2017) lay out, at times it may be that some treatment is applied, and the share of units receiving treatment may increase more in certain states than in other states. They call such a case "Fuzzy Differences-in-Differences", to distinguish it from the standard "sharp" design. In this case, the standard method is to calculate the treatment effect by calculating the difference-in-difference effect of the treatment variable on the outcome of interest, and then scaling by the impact of the treatment variable on the likelihood that one effectively receives treatment. Thus, to the degree that treatment does not increase completely, we will scale up the estimated effect to correct for the fact that only a sample of units were treated. We will return to such a design in the following section of these notes when discussing instrumental variables and "local average treatment effects."

One of the key results from (de Chaisemartin and D'Haultfoeuille, 2017) is to fully define the conditions under which this type of estimate will capture an unbiased average treatment effect, and to characterise the group for which this ATE will hold. In particular, they note that as well as a parallel trends assumption, we require assumptions that:

1. The ATE of units treated at multiple dates must be stable over time
2. When the share of treated units changes over time in control groups, the treatment effect for switchers in both treatment and control groups should be the same.

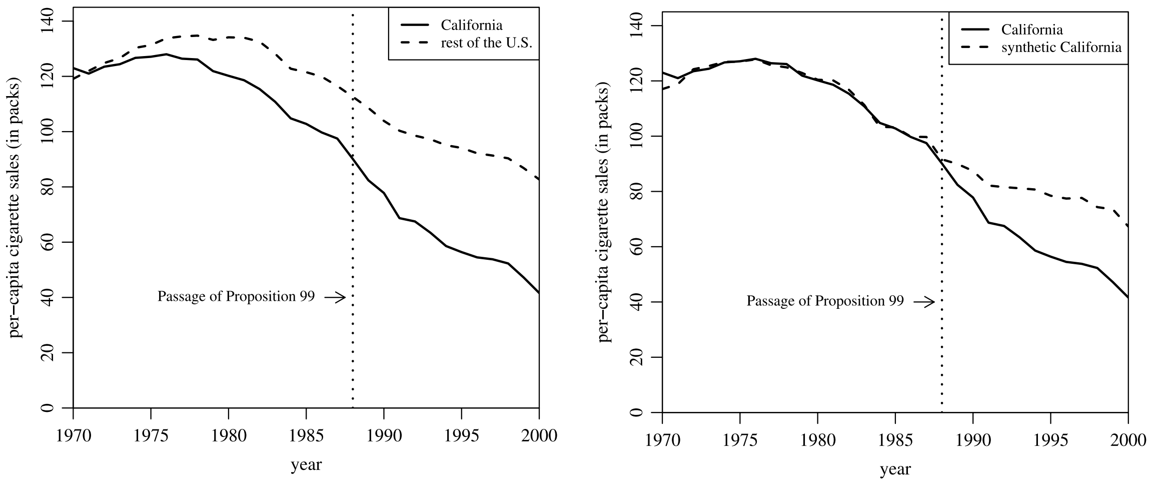
A second key contribution is that they propose alternative esitmators which can be used when the share of treated units in "control" groups is stable over time, and which no longer rely on these two additional assumptions.

2.3 Synthetic Control Methods

If, despite all our best efforts with differences-in-differences, event studies, or even triple differences, we do not manage to satisfy ourselves that parallel trends are met, fortunately all hope is not yet lost. One way to proceed even in the absence of parallel trends is by using synthetic control methods. These synthetic control methods aim to construct a "synthetic" (ie statistically produced) control unit for comparison with the true treatment unit. The synthetic control group is—similar to matching—formed using a subset of all potential controls, which are also known as donor units. These donor units are combined in a manner to track as closely as possible the trend in the true treatment group in the *pre*-reform periods. The logic behind the method is to form a comparison group as similar as possible to the control group considering only the pre-treatment data, and observe what happens once the treatment has taken place. If the synthetic control is a good match with the treatment group, all else constant, they should follow identical paths in the post-reform period. However, given that only the treatment group is affected by treatment receipt, we infer that any post-treament divergence in trends is due to the receipt of treatment itself. These methods, first discussed in Abadie and Gardeazabal (2003) were formalised in Abadie et al. (2010), whose exposition we follow below.

Graphically, figure 2.5 provides an example of the synthetic control process. In panel (a), we observe that outcomes in the treatment area (California) clearly diverge from those in the rest of the USA well before treatment occurs, and this divergence occurs in a way which violates the parallel trend assumption. However, in figure (b), we see that when a "synthetic control" is formed, this synthetic control group tracks the true outcomes in the treated area very well in the pre-reform period, however only diverges post-reform. It is this post-reform divergence that we interpret as our treatment effect.

Figure 2.5: Synthetic Controls and Raw Trends (Figures 1-2 from Abadie et al. (2010))



(a) No Synthetic Control

(b) Synthetic Control

The process of forming a synthetic control consists of assigning weights to potential control areas in such a manner to optimise pre-reform levels in the outcome variable. Following Abadie et al. (2010) we consider $J+1$ regions, one of which receives treatment, which we arbitrarily call region 1. The goal in synthetic control methods is to form a $J \times 1$ vector $\boldsymbol{W} = (w_2, \ldots, w_{J+1})'$ for which $w_j \geq 0 \ \forall j$, and $w_2 + \ldots + w_{J+1} = 1$. These weights are chosen so that they only use information prior to the reform of interest, and they ensure that all pre-reform average outcomes and controls are equalised between the treatment unit and the synthetic control unit. For example, in Abadie et al. (2010)'s example above, 5 of the potentially 49 donor states are given positive weights, while the remaining 44 states are given no weight, resulting in a near perfect fit in trends prior to the reform (figure 2.5b).

Assuming that these weights can be formed, this then suggests a reasonably simple way to calculate a treatment effect. We simply subtract from the *post*-reform outcome in the treatment state the weighted average of the *post*-reform controls in the synthetic control states:

$$\widehat{\alpha}_{1t} = Y_{1t} - \sum_{j=2}^{J+1} w_j^* Y_{jt}.$$

Note that in the above t refers only to post-reform periods. The existence of weights for estimation in particular requires that all pre-treatment outcomes and controls of interest in the treatment state are contained in a "convex hull" of the outcomes of the donor states, or that the values of the treatment state aren't universally higher or lower than those in all the donor states. We return to discuss what to do in the case this does not hold at the end of this section.

This idea captures the spirit of diff-in-diff methods, however rather than having to subtract the pre-reform difference from the post reform difference, the synthetic control ensures that the pre-reform difference is equal to zero. In order to actually implement this method, the question remains of how to calculate these weights. As Abadie et al. (2010) show, this can be treated as a problem of minimising the Euclidean norm (or roughly, the total average distance in many dimensions), as described below, where $\boldsymbol{V}$ is a semi-definite positive matrix:

$$\|\boldsymbol{X_1} - \boldsymbol{X_0}\boldsymbol{W}\|_{\boldsymbol{V}} = \sqrt{(\boldsymbol{X_1} - \boldsymbol{X_0}\boldsymbol{W})'\boldsymbol{V}(\boldsymbol{X_1} - \boldsymbol{X_0}\boldsymbol{W})}.$$

The full details of the weighting process, and indeed the estimator, are available in Abadie et al. (2010). What's more, the authors have made libraries available to implement this process in R, MATLAB and Stata, all available online.

Up until recently, where the treatment state had outcomes which were universally higher or universally lower than the donor states synthetic control methods could not be used. However, work from Doudchenko and Imbens (2016) extended synthetic control methods and loosened the estimation requirements. Principally, this allows for a constant different in levels between the treatment area and the synthetic controls. Doudchenko and Imbens (2016) document their updated methods using the same case as Abadie et al. (2010), and also a number of other applied examples.

Empirical Exercise 2: Suffrage and Child Survival

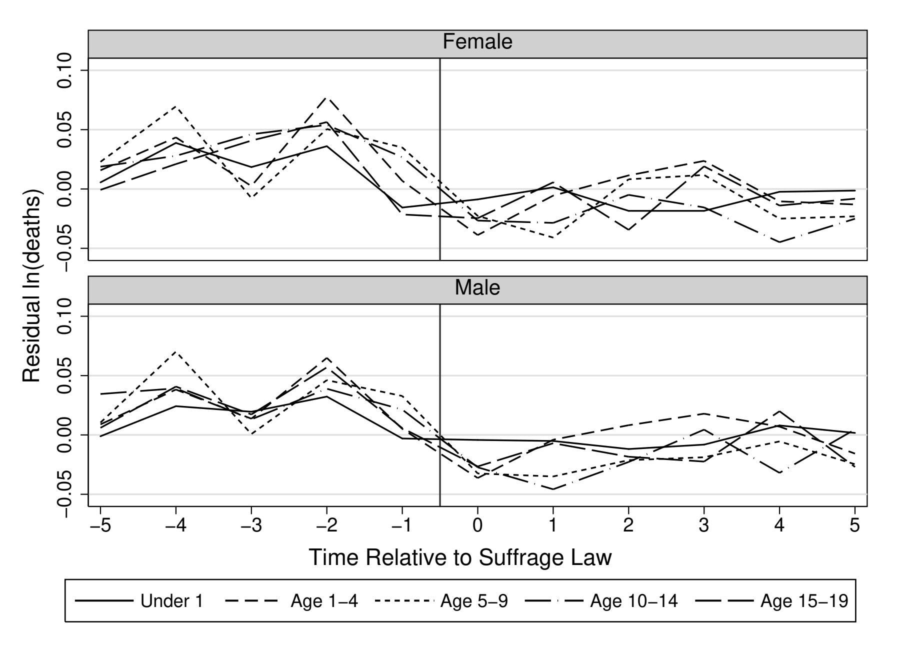
Instructions: In this exercise will examine the paper "Women's Suffrage, Political Responsiveness, and Child Survival in American History", by Miller (2008). We will first replicate the (flexible) difference-in-differences results examining the effect of women gaining the vote on child health outcomes using the dataset `Suffrage.dta` compiled from Grant Miller's website. We will then examine the importance of correct inference in a difference-in-difference framework, by examining various alternative standard error estimates both capturing and not-capturing the dependence of errors over time by state.

Questions:
(A) Replication of Principal Results

1. Replicate the results in table IV of the paper, following equation 1, as well as the notes to the table. [Note that in a small number of specifications you may find slightly different standard errors using this version of the data.]

2. Plot figure IV (and see below) from the paper using the same dataset for male and female mortality in each of the age groups displayed. Refer to the discussion on page 1306-1307 of Miller (2008) for details on calculations. This figure is based on average regression residuals for each year, and as you will likely remember, these regression residuals are calculated as $\hat{\varepsilon} = y - X\hat{\beta}$. These can be calculated in Stata following a regression by using the command `predict varname, resid`.

Replication of Figure IV of Miller (2008)



(B) Examination of Some Details Related to Inference For parts 1-3 of the below question, it is only necessary to report the p-values associated with each estimate (considering the null hypothesis that the coefficient on suffrage is equal to 0. In part 4, we are interested in the 95% confidence intervals of the estimate.

1. Replicate the results from table IV, however *without* using standard errors clustered by state.

2. Replicate the results from table IV using standard errors robust to heteroscedasticity

3. Re-estimate the results from table IV using wild bootstrap standard errors. This could be done using the user-written ado `boottest` which can be installed in Stata using `ssc install boottest`. If doing so, I suggest using the "noci" option of `boottest`.

4. Create a graph showing two sets of 95% confidence intervals for each estimate displayed in table IV: the first using clustered standard errors and the second using the uncorrected standard errors from point 1 above. Ensure to indicate where zero lies on the graph to determine which estimates are statistically distinguished from 0 at 95% in each case.

Chapter 3

Estimation with Local Manipulations: LATE and Regression Discontinuity

Required Readings
Imbens and Wooldridge (2009): Sections 6.3 and 6.4
Angrist and Pischke (2009): Chapter 4.1, 4.4, 4.5 (LATE) and Chapter 6 (RD)
Lee and Lemieux (2010): RD

Suggested Readings
Imbens and Angrist (1994)
Angrist et al. (2010)
Bharadwaj et al. (2013)
Clots-Figueras (2012)
Brollo and Troiano (2016)
Angrist and Lavy (1999)

In this section we will begin by returning to the relationship between what we have called *unconfoundedness* and the zero-conditional mean assumption that we used to define the exogeneity of our regressors in earlier econometrics courses when working with OLS. To do so, let's start with the *Rubin causal model*. Our workhorse example consists of potential outcomes,

$$y_{0i} = \mu_0 + \beta x_i + e_{0i} \tag{3.1}$$
$$y_{1i} = \mu_1 + \beta x_i + e_{1i} \tag{3.2}$$

and an assignment mechanism for W_i, which may depend on the values of X and e.

Given a set of observed variables (y_i, x_i, w_i), we can translate this into an estimable equation via the identity of the *switching regression*. But the 'right' way to write down this regression depends on what it is we are trying to estimate. Suppose first that we are interested in estimating the ATE. This is given by $\mu_1 - \mu_0$, the average difference between potential outcomes in the entire population. Writing

$$y_i = \mu_0 + \underbrace{(\mu_1 - \mu_0)}_{\tau^{ATE}} w_i + \beta x_i + \underbrace{(e_{1i} - e_{0i})w_i + e_{0i}}_{e_i^{ATE}}, \tag{3.3}$$

we can clearly see the requirement of exogeneity. We require w_i to be uncorrelated with the compound error term e_i^{ATE}. This requires unconfoundedness as we have defined it: w_i must be uncorrelated with *both* potential outcomes, y_{1i}, y_{0i}.[1]

Notice that if we were willing to *assume* that everyone had the same treatment effect, then $e_{1i} - e_{0i} = 0$, for all i, so in a constant effects model we can estimate the ATE even if we only have independence of w_i from e_{0i}. But if we are not willing to assume a constant effects model, then in general the ATT and the ATE will not coincide. If instead we are interested in estimating the ATT, then the expected value of $E[e_{1i} - e_{0i}|W_i = 1]$ is part of what we want to study. If the treated benefit more (or less) than the average member of the population, than this should be reflected in our estimate of the ATT. In this case let us write

$$y_i = \mu_0 + \underbrace{(\mu_1 - \mu_0) + (e_{1i} - e_{0i})}_{\tau^{ATT}} w_i + \beta x_i + \underbrace{e_{0i}}_{e_i^{ATT}}. \tag{3.7}$$

From this we can see that the exogeneity assumption required for regression to provide an unbiased estimate of the ATT is weaker than for the ATE. We require only that w_i is uncorrelated with e_{0i}, but *not* with e_{1i}. All of this leads us to the fact that unconfoundedness gives the zero conditional mean assumption that has traditionally been used to define exogeneity.

This is all well and good, but in the absence of a randomized, controlled trial, arguing for the assumption of unconfoundedness is often an uphill battle. We are therefore interested in what

[1] A brief description of why unconfoundedness satisfies this requirement is as follows. Unconfoundedness gives us (by definition) that $E[e_{1i}|w_i] = 0$ and that $E[e_{0i}|w_i] = 0$. Our challenge is to show that this implies the zero conditional mean assumption, namely, that $E[e_i^{ATE}|w_i] = 0$, where e_i^{ATE} is defined as in equation (3.3) as

$$e_i^{ATE} = (e_{1i} - e_{0i})w_i + e_{0i} = e_{1i}w_i - e_{0i}w_i + e_{0i}. \tag{3.4}$$

The expected value of the third term is zero by assumption, leaving us with the first two terms. We will show that $E[e_{1i}w_i|w_i] = 0$; the other follows by symmetry.

Take the case where $w_i = 1$. Then:

$$E[e_{1i}w_i|w_i = 1] = E[e_{1i}1|w_i = 1] = E[e_{1i}] = 0, \tag{3.5}$$

where the second equality follows from the unconfoundedness assumption. Alternatively when $w_i = 0$,

$$E[e_{1i}w_i|w_i = 0] = E[e_{1i}0|w_i = 0] = 0, \tag{3.6}$$

which completes the proof.

ways we can estimate the causal effects of a program under weaker assumptions. In what follows we will consider two cases where we can estimate a causal treatment effect *locally* (that is to say for some specific group), but not *globally*. We will first consider the case of instrumental variables and treatment effects, and then move on to regression discontinuity methods.

3.1 Instruments and the LATE

To understand the use of instrumental variables to estimate treatment effects, we return to our simplest case of potential outcomes without covariates:

$$\begin{aligned} y_{0i} &= \mu_0 + e_{0i} \\ y_{1i} &= \mu_1 + e_{1i}. \end{aligned} \tag{3.8}$$

We will begin by assuming *homogenous treatment effects*. Let $e_{0i} = e_{1i} = e_i$ for all individuals i. The resulting empirical specification is now

$$y_i = \mu_0 + \underbrace{(\mu_1 - \mu_0)}_{\tau} w_i + e_i. \tag{3.9}$$

If unconfoundedness holds, we can use OLS to estimate the parameter τ, which gives the ATE (equivalent to the ATT in this case). But what if unconfoundedness fails? Then the correlation between e_i, w_i means we have a (now familiar) endogeneity problem.

3.1.1 Homogeneous treatment effects with partial compliance: IV

In the case of homogeneous treatment effects, you are likely already familiar with one way of addressing this problem: instrumental variables. Suppose we have an instrument, z, that affects the likelihood of an individual receiving the treatment, w, but has no direct effect on the outcome of interest. Such an instrument will satisfy the *exclusion restriction* and *rank condition* required for standard instrumental variables estimation (Wooldridge, 2002, chapter 6).

The paradigmatic example of this is a randomized, controlled trial with imperfect compliance. Individuals may be assigned at random to treatment and control arms of the trial, but it is possible that some of those assigned to treatment may not undertaken the treatment, and some of those assigned to control arms may end up getting the treatment. In this case, so long as the initial assignment was truly random and has some power over which treatment people end up receiving, it can be used as an instrument. There are several ways to implement such an instrumental variables approach, which we examine below in turn.

(i) Two-stage least squares Two-stage least squares combines our causal model for the outcome,

$$y_i = \mu_0 + \tau w_i + e_i \tag{3.10}$$

with a first-stage regression that is a linear projection of the treatment on the instrument:

$$w_i = \gamma_0 + \gamma_z z_i + v_i. \tag{3.11}$$

Substituting the predicted values of w_i, $\hat{w}_i$, from the first-stage regression into the second stage regression gives

$$y_i = \mu_0 + \tau^{2SLS}\hat{w}_i + u_i. \tag{3.12}$$

where τ^{2SLS} consistently estimates the ATE. As usual, doing this in two stages by hand does not correct standard errors for the use of a constructed regressor, but these can be obtained directly by use of Stata's `ivregress` or related commands.

(ii) Indirect least squares It is also useful to understand that the 2SLS estimate can be reproduced from a pair of 'reduced-form' regressions. In particular, consider estimation of equation (3.11) together with the reduced form

$$y_i = \pi_0 + \pi_z z_i + \eta_i. \tag{3.13}$$

Now, recall the properties of the 2SLS estimator that τ is equal to the ratio of the covariances

$$\tau^{IV} = \frac{\text{cov}(y,z)}{\text{cov}(w,z)} \tag{3.14}$$

$$= \frac{\text{cov}(y,z)/\text{v}(z)}{\text{cov}(w,z)/\text{v}(z)}. \tag{3.15}$$

The second line follows just from dividing both numerator and denominator by the same quantity, the variance of z. This is helpful because the numerator and denominator are exactly what is estimated by the regression coefficients on z_i in the reduced-form and first-stage equations, respectively. That is, $\pi_z = \text{cov}(y,z)/\text{v}(z)$, and $\gamma_z = \text{cov}(w,z)/\text{v}(z)$. So, an indirect squares approach to estimating τ is to estimate the two reduced-form coefficients, and then take their ratio.

(iii) Wald estimator In the special case where our instrument is binary, equation (3.15) has a particularly useful interpretation. Notice that if z is binary, then the coefficient on this variable in the reduced-form regressions will give us the simple difference in means:

$$\pi_z = E[y|z=1] - E[y|z=0]$$
$$\gamma_z = E[w|z=1] - E[w|z=0].$$

Substituting these values into the ratio for indirect least squares (equation 3.15) gives the *Wald estimator*

$$\tau^{WALD} = \frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{E[W_i|Z_i = 1] - E[W_i|Z_i = 0]} \tag{3.16}$$

where τ estimates the ATE (=ATT, since we are still maintaining the assumption of homogeneous treatment effects). This is an application of a standard interpretation of instrumental variables to the case of a binary instrument; see Angrist and Pischke (2009) and Imbens and Wooldridge (2009) for discussion.

Once we relax the (strong!) assumption of homogeneous treatment effects, however, we can no longer interpret IV estimates as estimating 'the' treatment effect. *In fact, IV will not necessarily give us either the ATE or the ATT!*

3.1.2 Instrumental variables estimates under heterogeneous treatment effects

When treatment effects may be heterogeneous—and there is often little reason to rule this out *a priori*—and compliance with randomization into treatment is imperfect, the situation becomes considerably more complicated. It is now only under special conditions that we can estimate even the ATT (let alone the ATE).

In this context, in order to be able to interpret IV estimates as giving average treatment effect for *some* subpopulation, we will need stronger assumptions than are typically made in a homogeneous-effects IV world. This requires us to expand our potential outcomes notation, to be explicit about the effect of the instrument on treatment status and outcomes.

The possibility of noncompliance leads to an alternative measure of the treatment effect. Suppose we want to know what is the total benefit of our randomly assigned instrument. In many cases this may be the actual intervention: e.g., Z could be a conditional cash transfer program, and W could be schooling, Y a socio-economic outcome of interest.

Since our costs are associated with implementing Z, we may want to know the average benefit of those who receive $Z = 1$. This is the ITT:

Definition 1. *Intent-to-Treat effect*

The ITT is the expected value of the difference in outcome, Y, between the population randomly assigned to treatment status $W = 1$ (but who may not have ended up with that status) and those who were not:

$$ITT = E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]. \tag{3.17}$$

A useful result, due to Bloom (1984), relates the ITT to the ATT under the additional assumption that there is no *defiance*, that is, that Pr$[W_i = 1|Z_i = 0] = 0$:

$$ITT = ATT \times c, \tag{3.18}$$

where c is the *compliance rate*, $c =$ Pr$[W_i = 1|Z_i = 1]$. This follows intuitively from the independence of Z_i and potential outcomes (so that it is uncorrelated with Y_0).

3.1.3 IV for noncompliance and heterogeneous effects: the LATE Theorem

Under imperfect compliance, we have two *potential outcomes* in terms of W, for any given value of the instrument Z. For the two possible values of $Z_i \in \{0,1\}$, we define (W_{0i}, W_{1i}) as the corresponding potential outcomes in terms of *realised* treatment status. We can then write

$$W_i = W_{0i}(1 - Z_i) + W_{1i}(Z_i). \tag{3.19}$$

Notice also that the outcome variable may conceivably depend on on *both* treatment status *and* the value of the instrument. Let us denote by $Y_i(W, Z)$ the potential outcome for individual i with treatment status W and value of the instrument Z. So there are now *four* potential outcomes for each individual, associated with all possible combinations of W and Z.

The instrument, Z_i, will be *valid* if it satisfies the unconfoundedness (conditional mean independence) assumption with respect to the potential outcomes in Y and W. Formally, we will assume:

Assumption 4. *Independence*

$$(Y_i(1,1), Y_i(1,0), Y_i(0,1), Y_i(0,0), W_{1i}, W_{0i}) \perp\!\!\!\perp Z_i. \tag{3.20}$$

Independence alone does *not* guarantee that the causal channel through which the instrument affects outcomes is restricted to the treatment under study. For this reason, we add the standard exclusion restriction:

Assumption 5. *Exclusion restriction*

$$Y_i(w, 0) = Y_i(w, 1) \equiv Y_{wi} \tag{3.21}$$

for $w = 0, 1$.

An individual's treatment status fully determines the value of their outcome, in the sense that the instrument has no direct effects.

A standard requirement for instrumental variables, including this case, is one of *power*. When IV was introduced, we required the instrument to be partially correlated with the endogenous variable, conditional on the exogenous, included regressors (Wooldridge, 2002, ch. 5).

Assumption 6. *First stage*

$$E[W_{1i} - W_{0i}] \neq 0. \tag{3.22}$$

Notice that this is a statement about the expected value for the population as a whole. It does not guarantee that any individual is 'moved' by the instrument to change their treatment status. It does not even guarantee that all individuals are 'moved' in the same direction: some may be induced by the instrument to take up treatment, whereas they otherwise would not have done so, while others may be induced by the instrument not to take up treatment, whereas they otherwise would have done so.

For this reason, interpretation of an IV regression as the treatment effect for some subpopulation requires something stronger than first-stage power alone. In particular, we require that *all individuals in the population* are uniformly more (or less) likely to be treated when they have $Z_i = 1$.

Assumption 7. *Monotonicity*

$W_{1i} \geq W_{0i}, \forall i.$

Notice that if the instrument takes the form of a discouragement from taking up the treatment, we can always define a new variable $Z_i' = (1 - Z_i)$, which will satisfy monotonicity as defined above.

Under these four conditions, instrumental variables estimation will give us a *local average treatment effect*—an average treatment effect for a specific subpopulation. The LATE Theorem (Angrist and Pischke, 2009, p. 155) gives us…

Theorem 2. *The LATE Theorem*

Let $y_i = \mu_0 + \tau_i w_i + e_i$, and let $w_i = \gamma_0 + \gamma_{zi} z_i + \eta_i$. Let assumptions 1 - 4 hold. Then

$$\begin{aligned}\frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{E[W_i|Z_i = 1] - E[W_i|Z_i = 0]} &= E[Y_{1i} - Y_{0i}|W_{1i} > W_{0i}] \\ &= E[\tau_i|\gamma_{zi} > 0]\end{aligned}$$

See Angrist and Pischke for the proof, which is not reproduced here.

3.1.4 LATE and the compliant subpopulation

The LATE theorem tells us that the Wald/IV estimator provides an unbiased estimate of treatment effects for some subpopulation—the subpopulation for whom $W_{1i} \neq W_{0i}$. Who are these people?

The answer, unfortunately, depends on the instrument that we are using, and its ability to affect the eventual treatment status of individuals in the sample. Relative to a given instrument, we can categorize individuals in four groups. These are listed in table 3.1). Notice here, that the assumption of monotonicity rules out the existence of *defiers*.

Table 3.1: Compliance types

Group	Definition	Words
Compliers:	$W_{1i} = 1, W_{0i} = 0$	Participate when assigned to participate, don't participate when not assigned to participate
Never-takers:	$W_{1i} = 0, W_{0i} = 0$	Never participate, whether assigned to or not
Always-takers:	$W_{1i} = 1, W_{0i} = 1$	Always participate, whether assigned to or not
Defiers:	$W_{1i} = 0, W_{0i} = 1$	Participate when assigned not to participate, don't participate when assigned to participate

Our estimates of the treatment effect will be entirely driven by the compliers. With IV we estimate a *Local Average Treatment Effect*: the average treatment effect on the compliant subpopulation. This implies that IV is not informative for always takers and for never takers, as the intrument has no power to shift the treatment status for these groups. As an aside, this is somewhat analagous to fixed effects models in panels, where estimates are driven only by units who 'change' within the panel.

For this reason we may want to be able to say something about who exactly these compliers are. Under monotonicity, the size of the compliant sub-population is given by the first stage of our IV estimation (Angrist and Pischke, 2009, p. 167):

$$\begin{aligned}\Pr[W_{1i} > W_{0i}] &= E[W_{1i} - W_{i0}] \\ &= E[W_{1i}] - E[W_{0i}] \\ &= E[W_i|Z_i = 1] - E[W_i|Z_i = 0] \qquad (3.23)\end{aligned}$$

where the last line makes use of the independence assumption. We can use this to determine the fraction of the treated who are compliers (Angrist and Pischke, 2009, p. 168). If a high proportion of the treated are compliers, we can feel relatively confident about the representativeness of the estimated treatment effect.

Different instruments will have different populations of compliers, and so different LATEs. This insight has important lessons for tests used elsewhere for the validity of instruments. If treatment effects are heterogeneous, and we estimate very different effects using two different instruments, we may not be able to tell whether this is due to heterogeneity in treatment effects or due to the invalidity of one of the instruments.

Can we say anything about the characterisitcs of compliers? While the above suggests that it is simple to know what proportion of observations are compliers, we likely would like to be able to say more about this group to better interpret the LATE. Of course, we cannot simply "look at" the compliers and summarise their observations, given that one's status as a complier is unobservable. However, we can still say *something* about the relative frequency of characteristics among compliers, allowing us to respond to questions such as: "are the compliers more likely to have a secondary education than the general population?" or any such question relating to observed characteristics. To see this, note that (from Angrist and Pischke (2009, p. 171)) for a binary variable x_{1i}:

$$\frac{\text{Pr}[x_{1i}=1|W_{1i}>W_{0i}]}{\text{Pr}[x_{ii}=1]}=\frac{\text{Pr}[W_{1i}>W_{0i}|x_{1i}=1]}{\text{Pr}[W_{1i}>W_{0i}]}=\frac{E[W_i|Z_i=1,x_{1i}=1]-E[W_i|Z_i=0,x_{1i}=1]}{E[W_i|Z_i=1]-E[W_i|Z_i=0]}.$$

That is to say, if we would like to know how much more/less likely the population of compliers is with characteristic $x_{i1}=1$ versus the whole population (the left-hand term), we simply need to compare the first stage for this group, with the first stage for the whole population. We can follow a procedure of this type to examine the distribution of any variables of interest.

Treatments with Multiple Levels So far, we have been considering the case where W_i, the endogenous (treatment) variable of interest is a binary variable. In this case when we estimate a LATE, although the parameter is "local" to some specific group, it is clear that this parameter refers to the impact of a shift from 0 to 1 in the binary variable W_i. However, how does the interpretation of the LATE change when we consider a multi-level treatment variable? For example, what if the treatment variable of interest is years of schooling, or total fertility in a family? The response to this question turns out to require thinking not just about *who* the instrument causes to shift behaviour, but also thinking about *at what margin* of the dependent variable the instrument induces shifts.

For example, consider a case where we wish to examine the impact of fertility on child educational outcomes. Here a frequently used instrument is the twin birth instrument. In particular, let's consider a twin birth at birth order 3. This could cause families to move from having had 3 children without the twin, to four children with the twin, but similarly, could also cause higher margin shifts in fertility, eg a family that would have had 4 births now has 5 births. In the case of a multi-leveled treatment variable, as the instrument can cause shifts at multiple margins of the treatment variable we can no longer talk of a single class of complier. Here, Angrist and Imbens (1995) show that the interpretation of the parameter is now in terms of the "Average Causal Response" function, or in terms of the entire shift of the distribution of the endogenous variable of interest caused by receipt of the instrument. To do this, they define the Average Causal Response (ACR) function as follows:

$$\begin{aligned}\frac{E[Y_i|Z_i=1]-E[Y_i|Z_i=0]}{E[S_i|Z_i=1]-E[S_i|Z_i=0]} &= \sum_{s=1}^{S}\omega_s E[Y_{si}-Y_{s-1,i}|s_{1i}\geq s>s_{0i}] \\ \text{where} \qquad \omega_s &= \frac{P[s_{1i}\geq s>s_{0i}]}{\sum_{j=1}^{S}P[s_{1i}\geq j>s_{0i}]}\end{aligned}$$

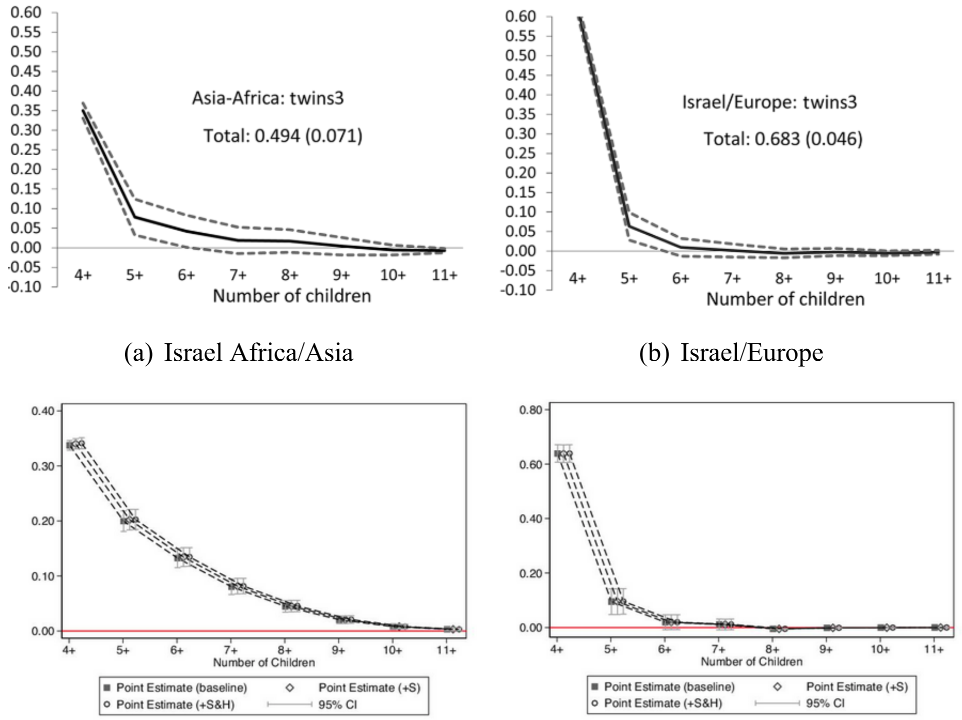
Here, we use S_i to refer to the multi-leveled treatment variable of interest. Note that the quantity on the left-hand side of the first line is the Wald estimate that would come out of our IV model. Thus, the ACR theorem states that we can interpret the IV (Wald) estimate with a multi-levelled treatment as a weighted average of the effect of compliers who are shifted by the instrument to move from $s-1$ to s at each point of the distribution of S, where the weights are given by the probability that the instrument shifts the distribution of S at this point. In this case, this is another reason why LATEs from different (valid) instruments may not be the same if the instrument traces out different shifts in this ACR function.

These weights can perhaps be iilustrated most simply with an example. Consider the case mentioned briefly above, where we wish to estimate the impact of family fertility (a multi-valued treatment) on children's educational outcomes. In the below panels, we show plots of the ACR function of shifts in fertility induced by a twin birth. The top two panels are based on different samples in census data in Israel (from Angrist et al. (2010)), and the bottom two panels are based on a developing country sample and a sample from the USA from Bhalotra and Clarke (2019a). While in all cases a twin at birth order three causes the biggest shift in fertility when considering families exceeding 4 children, it also has higher order impacts, up to as much as 9+ births in a developing country sample.[2] Note that here, even using the same instrument, the LATE will be interpreted quite differently depending on the context examined.

What Changes if we add Controls? In general when discussing the LATE, the Wald estimate is considered as a starting point, however in practice the assumptions leading to consistency in IV estimates may only hold conditional on covariates. For example, the frequently used twin instrument is thought to require (at least) controls for maternal age, given that the likelihood of a twin birth increases as women become older due to hormonal changes (see for example discussion in Bhalotra and Clarke (2019b)). In practice, this implies that we are replacing the independence assumption (3.20) with a conditional indendence version:

[2] These shifts can be rationalized if one thinks about issues such as access to contraceptive measures and labour market opportunites, among other things.

Figure 3.1: Average Causal Response Functions for Twin Birth



(a) Israel Africa/Asia

(b) Israel/Europe

(c) Developing Countries

(d) USA

Assumption 8. *Conditional Independence*

$$(Y_i(1,1), Y_i(1,0), Y_i(0,1), Y_i(0,0), W_{1i}, W_{0i}) \perp\!\!\!\perp Z_i|X_i. \tag{3.24}$$

To see what we estimate if we run 2SLS with controls, Angrist and Pischke (2009) introduce the following notation:

$$\lambda(X_i) \equiv E[Y_{1i} - Y_{0i}|X_i, D_{1i} > D_{0i}],$$

where $\lambda(X_i)$ refers to the treatment effect for each possible value of X_i. For example, consider a simple case where the only covariate is maternal age. In this case, we would have a single $\lambda(X_i)$ value for each maternal age observed in the population. This is what Angrist and Pischke (2009) refer to as a saturated model—a model where all possible levels of the covariates are included as a series of dummy variables. In this case, they show that the treatment effect estimated with a full set of saturated covariates is:

$$\tau = E[\omega(X_i)\lambda(X_i)], \qquad \text{where} \qquad \omega(X_i) = \frac{V(E[W_i|X_i, Z_i]|X_i)}{E[V(E[W_i|X_i, Z_i]|X_i)]},$$

where $V(E[W_i|X_i, Z_i]|X_i) = E\{E[W_i|X_i, Z_i](E[W_i|X_i, Z_i]-E[W_i|X_i])|X_i\}$. Thus, in words, the 2SLS estimand is a weighted average of each covariate-specific LATE, where the weights are given by $\omega(X_i)$. These weights place more emphasis on groups for which the instrument creates more valuation in the fitted values of the first stage: ie groups where the instrument produces more variation in treatment conditional on covariates. Angrist and Pischke refer to this as the "Saturate and Weight" theorem, though it is important to note that there may be times when we wish to work with more parsimonious models, for example models where continuous variables enter linearly rather than as a series of fully saturated dummies. In this case, there is a result based on work from Abadie (2003) which states that *for compliers* the treatment versus control comparison conditional on X_i is equal to LATE conditional on X_i. However, in practice, the challenge is that we do not know who the compliers are, so we cannot estimate these LATEs directly. Abadie (2003) introduced what is now known as "Abadie's Kappa", which allows us to "find" compliers, and hence estimate this LATE directly in the complier group. This κ term is useful for a number of reasons, however goes somewhat beyond the scope of these lectures. To read more, refer to Angrist and Pischke (2009, pp. 178–180) and references there-in. An important takeaway from this is that to the degree that the probability that Z_i is "switched-on" is approximately linear in X_i, the 2SLS estimand will approximately estimate the conditional LATE for compliers.

3.1.5 Some Closing Points on the LATE

The Local Average Treatment Effect is what is delivered from a binary instrumental variable with heterogeneity, however it is likely not the quantity that we are most intersted in estimating for policy reasons. While the precise nature of the compliers will depend on each IV, policy relevant quantities are likely based on entirely different criteria, such as the impact on the entire population, or the impact on some particular targeted group. There is a robust discussion of the utility of the LATE, focusing on (among other things) the relative importance of the (good) internal validity of the estimates under the maintained assumptions, versus the parameter's use in an external population. Much has been written here. A useful (more positive) take of the LATE is provided by Imbens (2010) in a paper entitled "Better LATE than Nothing…". A more critical view is provided by Deaton (2009), a small portion of which is provided below.

> "*The LATE may, or may not, be a parameter of interest to the World Bank or the Chinese government and in general, there is no reason to suppose that it will be. For example, the parameter estimated will typically not be the average poverty reduction effect over the designated cities, nor the average effect over all cities.*
>
> *I find it hard to make any sense of the LATE. We are unlikely to learn much about the processes at work if we refuse to say anything about what determines θ; heterogeneity is not a technical problem calling for an econometric solution, but is a reflection of the fact that we have not started on our proper business, which is trying to understand what is going on. Of course, if we are as skeptical of the ability of economic theory to deliver useful models as are many applied economists today, the ability to avoid modeling can be seen as an advantage, though it should not*

> *be a surprise when such an approach provides answers that are hard to interpret.*" Deaton (2009, pp. 9–10).

Later in these notes we will return to other quantities estimated based on similar types of models when returning to discuss heterogeneity in more detail. Regardless of your own opinion of the use of LATE, it is important to understand exactly what is being estimated in these models, given their frequency of appearance in papers in economics.

3.2 Regression Discontinuity Designs

3.2.1 An Introduction to RDDs

We may not always be willing to assume that the relevant unobservables driving both potential outcomes and treatment assignment are time-invariant as was the case in differnce-in-differences style models we have studied previously. An alternative is to assume that unconfoundedness holds *locally*, i.e., only in a small neighborhood defined by an observable correlate of selection.

For example, if we were interested in examining the effect of different types of politicians on the outcomes in their constituencies, we would be very hard-pressed to make the claim that politicians are randomly assigned to localities, given that they are explicitly chosen (elected) by constituents! However, in a reasonably tight margin, we may be willing to assume that the difference between a politician gaining slightly more than a majority of the vote or gaining slightly less than the majority is largely random. In the limit, the difference between 50% and +1 vote and 50% -1 vote is extremely small, and plausibly unrelated to potential outcomes. However, the final result of both elections is radically different. In the first case, the assignment mechanism implies that the consituency recieves treatment (the politician in question), while in the second case the constituency does not receive tretment. Such **local unconfoundedness** type assumptions are at the heart of the regression discontinuity approach. It turns out that such arbitrary discontinuities are not infrequent in practice, as often formal decision rules are needed where various individuals seek access to limited spots. For example, discontinuities are encountered in educational admissions based on test scores, diagnostic decisions to define medical care, access to means tested public programs, and a whole host of other circumstances.

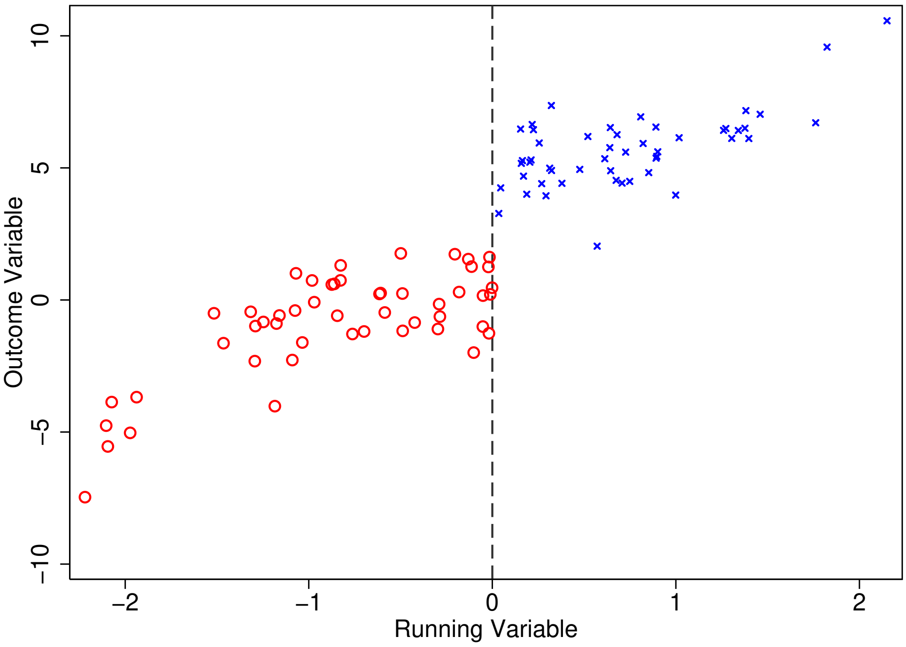
Notice that when assignment of treatment status varies according to strict rules along a single observable dimension, x, then we have a special problem for matching methods. On the one hand, enforcement of the rule means that the assumption of common support will be violated—we will inevitably rely on some kind of extrapolation. On the other hand, such a rule itself provides us with the ability to be confident about the process of selection into the program (particularly when it is sharply enforced). There may be no problem of selection on unobservables in this case; our primary concern is now allowing an appropriate functional form for the direct effect of the selection criterion x on the outcome of interest.

Following Lee (2008), suppose that treatment is assigned to all individuals with x greater than or equal to cutoff κ. The variable x (vote share in the above example) has a direct effect on outcomes of interest, such as corruption. If we are willing to assume that this effect is linear, then we can use regression methods to estimate:

$$y_i = \beta_0 + \beta_x x_i + \tau w_i + u_i \tag{3.25}$$

where τ will give us the ATE. If the rule is perfectly enforced, then conditional on x there is no correlation between w_i and u_i (i.e., conditional mean independence will hold), so τ is an unbiased estimate. But in order to do this, we must be very sure that we have the functional form right for the relationship between x and potential outcomes.

Figure 3.2: Strict Regression Discontinuity Design

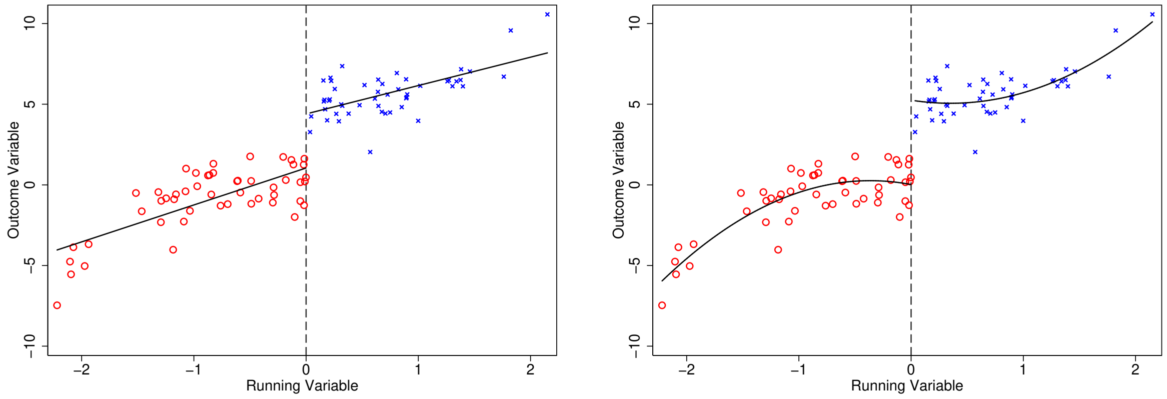


Consequently, we may want to be more cautious in extrapolating a linear relationship between x and y. This is illustrated in Figure 3.2, where a simple plot of the data suggests that extrapolating a linear functional form for the relationship between x and potential outcomes may be inappropriate (in fact the true DGP in this simulated example is a cubic function).[3] This is illustrated in Figure 3.3. In panel (a), the “discontinuity” observed between the two linear

[3] Figure 3.2 presetns a regression discontinuity setting with a perfectly enforced eligibility rule (at $x = 0$). Treated individuals are denoted by the small blue x, untreated by the red o. The DGP of y is $y = 0.6x3 + 5w + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0,1)$. Linear regression of y on x and w gives β_x = 2.08(0.22) and τ = 3.29(0.42).

predictions at point 0 is considerably smaller than the discontinuity observed when a quadratic fit is considered in panel (b). Here, if a linear fit were considered, extrapolation leads us astray: in this case, it leads us to dramatically underestimate the true treatment effect. Extrapolation is required in particular here precisely because the clean enforcement of the eligibility rule creates a situation of zero overlap. We never observe y_0 for $x > \kappa$, for example. Drawing on a similar logic to propensity score matching, we can relax functional form assumptions by comparing outcomes *only* among individuals who are in a neighborhood of x suitably close to the boundary.

Figure 3.3: Regression Discontinuity and The Running Variable



(a) Linear Fit (b) Quadratic Fit

Local unconfoundedness: We now make a less stringent assumption about (non-)selection on unobservables: the unconfoundedness needs only hold locally, in a neighborhood around κ. As Lee and Lemieux (2010) argue, even when agents can exert control over the forcing variable x, if that control is imperfect then the realization of whether x is above of below the cutoff κ, for agents very close to κ, is likely to be driven largely by chance:

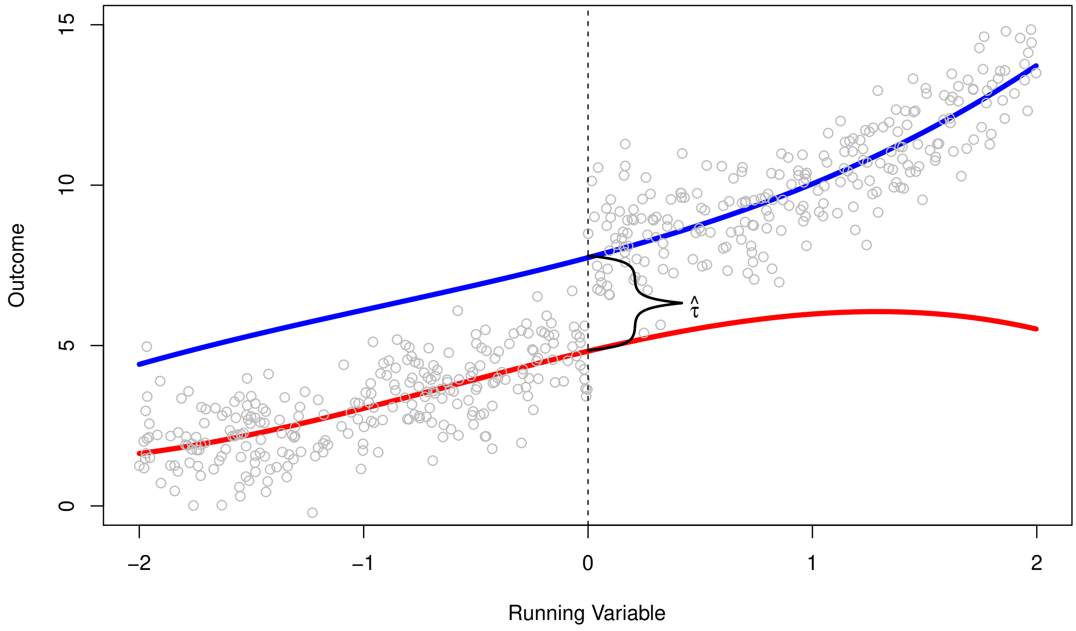
$$\lim_{x\downarrow\kappa} E(\varepsilon_i|x > \kappa) = \lim_{x\uparrow\kappa} E(\varepsilon_i|x < \kappa).$$

If local unconfoundedness holds, this then leads to our estimate of the effect of treatment:

$$\begin{aligned} \tau &= \lim_{x\downarrow\kappa} E(Y_i|x > \kappa) - \lim_{x\uparrow\kappa} E(Y_i|x < \kappa) \\ &= \lim_{x\downarrow\kappa} E(Y_{1i}|x > \kappa) - \lim_{x\uparrow\kappa} E(Y_{0i}|x < \kappa) \qquad (3.26) \\ &= E[Y_{1i} - Y_{0i}|x = \kappa] \end{aligned}$$

In general, what we estimate in a regression discontinuity is the average treatment effect for observations with x approximately equal to κ. When treatment effects are heterogeneous, this will not be either the ATE or the ATT, but rather the $ATE(\kappa)$. Of course, there is nothing that will imply that this treatment effect will tell us anything about treatment effects at other points

Figure 3.4: Regression Discontinuity and Heterogeneity over the Distribution of x



of the running variable — see for example figure 3.4 where $\tau(\kappa)$ is relatively uninformative for τ at certain other points of the support of the running variable.

The closer the neighborhood around κ we use for estimation, the less of an effect our assumptions about the functional form for x will have. But it is common to use a flexible or nonparametric approach for the relationship between x and y_i to avoid making assumptions about functional form in any case. These are described in section 3.2.3 below.

3.2.2 Regression Discontinuity Designs

Sharp Design

The prototypical RD design is a "sharp design", where the discontinuity implies a concrete change in treatment status at the threshold κ. In this case, all individuals who are located below the threshold value are assigned to the control group, and all individuals who are located above the threshold value are assigned to the treatment group. This allows to write a very simple model for treatment assignment, which is that:

$$w_i = \mathbb{1}\{x_i \geq \tau\}. \tag{3.27}$$

A clear example of this could be an election between two candidates (say a left-wing versus a right-wing mayor). If our treatment is that a county is assigned to a left-wing mayor, we know that this will only be observed if in this county the left-wing mayor receives at least 50%+1 vote, while if the candidate receives 50%-1 vote, the municipality will be assiged to the "treatment" group. The important thing here is that there is absolutely no discretion: if a majority of votes is received a candidate is chosen, whereas if a majority is not received, a candidate is not chosen.

Given the assignment mechansim described in equation 3.27, the impact of being assigned to treatment can be isolated easily given that the variable of interest jumps from 0 to 1 precisely when moving across point κ. In this case, we can estimate the effect of assignment to w_i following equation 3.26. When comparing average outcomes of y at points just below τ, with average outcomes of y at points just above τ we gain an esimate of the impacts of treatment shifting from 0 to 1, holding all else constant (save for the very small movement in the running variable).

Fuzzy Design

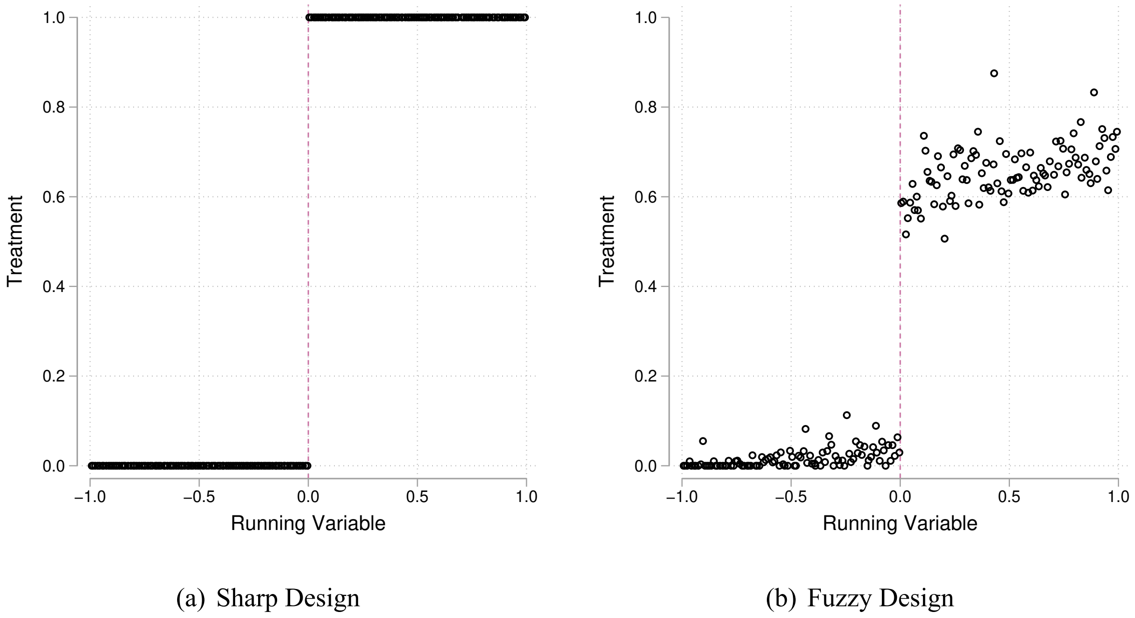
In the "sharp" regression discontinuity design examined so far, the probability of receiving treatment jumps deterministically from zero to one at the cut-off. Such is the case, for example, with simple majority elections, where crossing the threshold of the vote majority automatically results in a candidate being elected. Perhaps even more common than pure regression discontinuities are situations in which the probability of treatment jumps at the cut-off, but not deterministically. In these cases, not everyone above the cutoff is treated, and not everyone below the cutoff is untreated. Nevertheless, there is *some* local manipulation which ocurrs at this point, and which can be used for identification of a treatment effect. Essentially, now rather than the likelihood of treatment jumping by one at the cut-off, we observe:

$$\lim_{x\downarrow\kappa} Pr(w_i|x>\kappa) \neq \lim_{x\uparrow\kappa} Pr(w_i|x<\kappa). \tag{3.28}$$

Graphically, the difference in these designs can be obeserved in Figure 3.5. In the case of sharp designs, plotted on the right-hand side, there is no discretion in the application of the assignment rule. To the right of the threshold, no individual receives treatment, and to the left of the threshold, everybody receives treatment. However, in the case of the Fuzzy design, treatment is discretional. While in Fuzzy designs, the threshold *does* have an impact and does shift behaviour, this shift is not absolute, with certain individuals either opting into treatment below the cutpoint or opting out of treatment above the cutpoint, in which case the observed change does not jump sharply from 0 to 1, but rather in a "fuzzy" way from some value greater than or equal to 0, to some value less than or equal to one.

For example, Ozier (2011) uses a cutoff (eligibility) rule in primary exam scores to estimate the impact of secondary education in Kenya; not everyone who gets a score above the threshold attends secondary school, but at least some do. In such cases, instrumental variable methods

Figure 3.5: Fuzzy versus Sharp RDD Designs



(a) Sharp Design

(b) Fuzzy Design

may be used: the discontinuity may be thought of as a valid instrument for treatment in the neighborhood of the discontinuity. This is an interesting example of the LATE framework laid out above: the cut-off (treatment) provides a case of imperfect compliance. Now, rather than simply estimating the difference between those just above and just below the cut-off (as was the case in a sharp RD and equation 3.26), the effect must be weighted by the probability that those who cross the threshold are convinced to opt for treatment[4]:

$$\tau^F = \frac{\lim_{x\downarrow\kappa} E(Y_i|x > \kappa) - \lim_{x\uparrow\kappa} E(Y_i|x < \kappa)}{\lim_{x\downarrow\kappa} E(W_i|x > \kappa) - \lim_{x\uparrow\kappa} E(W_i|x < \kappa)}. \tag{3.29}$$

This is the well known Wald estimator. As in section 3.1.4, it allows us to estimate a treatment effect, but this treatment effect holds *only* for the subpopulation of compliers. In this case, compliers are the units who would get the treatment if the cutoff were at κ or above, but they would not get the treatment if the cutoff were lower than κ. In the Ozier (2011) example, they are those students who would go on to secondary if they achieve a score above the cut-off in the Kenyan Certificate of Primary Education, however would leave school if they do not achieve a score over the minimum cut-off.

[4] It is worth noting then, that as the denominator (likelihood of treatment given that the threshold is crossed) approaches 1, the fuzzy regression formula converges on the sharp RD formula displayed in 3.26. This is always the case with LATE, where as the instrument becomes perfectly binding, the IV estimate approaches the reduced form estimate.

3.2.3 Estimation and Inference with RD

Global vs Local Methods

Practical concerns when it comes to estimating parameters in regression discontinuity stem from the fact that we must adequately capture the relationship between the running variable and the dependent variable itself. If we fail to properly capture this relationship, we may incorrectly infer that this relationship is due to the discontinuity, κ rather than simply movements away from the discontinuity x.

There are two broad ways to deal with the issue of the relationship between the running variable and the outcome of interest. The first—parametric methods—consist of trying to adequately model the relationship between y and x over the entire range of data. The second—non-parametric methods—consist of limiting analysis to a short interval optimally chosen to be close to the cut-off (a distance known as the bandwidth), and then simply fitting a linear trend on each side.

Parametric methods These methods approach regression discontinuity as a problem of fitting a correctly-specified functional form to model the relationship between the running variable and the outcome variable on each side of the cut-off. Thus, the name "parametric methods", as we wish to correctly parametrize the relationship between x and y to thus isolate the effect of jumps in w at the threshold κ. These methods, also sometimes known as the global polynomial approach, then infer that the effect of receiving the treatment is the difference between each function as it approaches the discontinuity from each direction.

The global polynomial approach is straightforward to implement (Lee and Lemieux, 2010). It amounts to a regression of the form (here a second-order polynomial):

$$\begin{aligned} y_i &= \mu_0 + (\mu_1 - \mu_0)T_i + \beta_1^+ T_i(x_i - \kappa) + \beta_1^-(1 - T_i)(x_i - \kappa) \\ &\quad + \beta_2^+ T_i(x_i - \kappa)^2 + \beta_2^-(1 - T_i)(x_i - \kappa)^2 \end{aligned}$$

Notice that the polynomial is centered at the cutoff point and the polynomial can take a different shape on either side of the cutoff. These address potential non-linearity illustrated in Figure 3.2. Here, the estimates $\{\beta_1^+, \beta_1^-, \beta_2^+, \beta_2^-\}$ are designed to (adequately?) capture the relationship between x and y, while the treatment effect of interest is given by the remaining discontinuity at treatment T_i, which in our model is captured by $\mu_1 - \mu_0$.

The parametric approach thus reduces to correctly specifying these global polynomials. While the above specification suggests a cuadratic relationship, there is nothing (computationally at least) stopping us from using a cubic or even cuartic polynomial. The outstanding issue is then the choice of order of polynomial. One approach, described by Lee and Lemieux (2010), include choosing the model that minimizes the Akaike information criterion (AIC):

$$AIC = Nln(\widehat{\sigma}^2) + 2p$$

where $\widehat{\sigma}^2$ is the Mean Squared Error, and p is the number of parameters. An alternative is to include dummy variables for a number of bins, alongside the polynomial, and test for the joint significance of bin dummies. The latter is also useful as a form of falsification test: we might worry if there were discontinuities in the outcome variable at thresholds other than the cutoff we are using for analysis.

However, more generally, we should ask ourselves *why should we use all the data for inference if we are explicitly making a local identification argument?* Surely, if we are using data over a larger range of x values, we should be more concerned that the "local unconfoundedness" assumption becomes more and more unbelievable, and the marginal benefit of adding data very far from the discontinuity is highly questionable. These concerns are precisely why parametric approaches are rarely appropriate, and generally should not be used. In practice, regression discontinuity applications focus on local methods, where considerable attention is paid to the concern of how to determine the optimal analysis window.

Non-parametric methods then take the more logical approach of focusing only on a small sample of the data with a value of x that puts it very close to the cut-off point. By doing so, we line up the theory which states that falling on either side of the cut-off is locally random, with the practice of focusing on areas local to the cut-off.

Local Polynomial Methods

The idea behind local polynomial methods is that—in line with identifying assumptions—we will focus our attention on areas "local" to the cut-off. We will then parametrically control for x within this local area only, discarding observations which are too far from the cut-off to merit consideration. We call the interval around the cutoff that is used for estimation the **bandwidth**, generally denoted h. The limiting argument above in (3.26) hints at a key feature of the asymptotic argument that underlies the RD approach (Lee and Lemieux, 2010): the bandwidth should be as small as the sample allows. There are two main reasons for why this is advantageous. First, the bigger the bandwidth that we use, the more important it is to correctly specify the functional form for the relationship between the running variable, x, and potential outcomes. As the bandwidth shrinks, there is less and less variation in x in the sample being used for estimation, and so the scope for x to bias estimates of the treatment effect is reduced. Second, if x is chosen by agents under study, but without perfect control, then agents with very similar x values who end up on opposite sides of the cutoff are likely to have made similar choices. The reason that they end up on either side of the cutoff is largely chance. On the other hand, agents very far from the cutoff may have made different choices about x. Those differences may be too big to be likely to be explained by imperfect control of x. And if choice of x is determined with (even partial) knowledge of potential outcomes, then larger bandwidths introduce a source of bias.

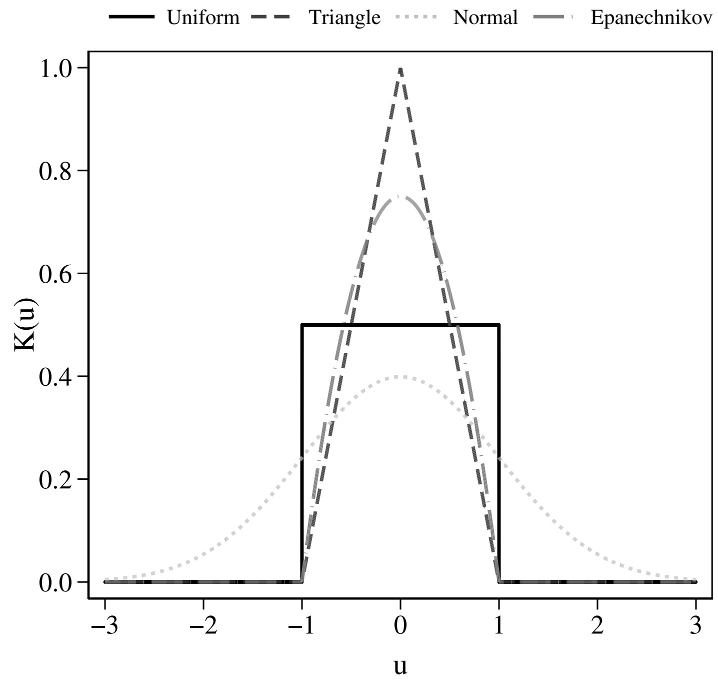
As laid out in Cattaneo and Titiunik (2022), the local polynomial approach requires four steps. These are:

1. Selecting the local polynomial order and kernal weighting function
2. Given these choices, determining a bandwidth h for estimation
3. Combine the choices from 1 and 2 with a standard least squares method for estimation
4. Conduct valid statistical inference

Selection of Polynomial Ordering Even when focusing on "local" windows around the cutoff point, it is necessary to control for relationships between x and y which may partially confound the RD estimate at the cut-point. In local polynomial methods, these controls consist of separate polynomial controls of x on either side of the cut-off. For example, a first order polynomial ($p = 1$) consists of capturing a separate linear relationship on either side of the cut-point, while a quadratic polynomial, $p = 2$, consists of including controls for x and x^2, with separate parmeters on either side of the cut-off. In practice, the standard recommendation is to use local linear methods, setting $p = 1$ (Cattaneo and Titiunik, 2022). This picks up recommendations to avoid higher-order polynomials, which have been made, for example by Gelman and Imbens (2019) who caution against using polynomials greater than order 2 to capture regression discontinuity effects, focusing on this practice in the global polynomial approach described previously. The point from Gelman and Imbens (2019) is that estimates using higher order polynomials may be misleading, and potentially harmful to estimated effects given that they may give unreasonable weight to values which are far from the cut-off in fitting polynomials, and may be very jumpy close to the cut-off point with important implications for the estimated parameters. Their preference is to focus on local linear regression discontinuity, or polynomials only up to quadratics (once again in a local setting) to optimally capture effects of the running variable. Recent work from Pei et al. (2021) suggest that in some *local* settings, higher polynomials may actually behave reasonably well, and propose an optimal (mean squared error minimizing) procedure to select the degree for local polynomials.

Selection of a Kernel A separate consideration relates to how to weight observations local to the cut-off. This is defined using a kernel density, generally referred to as $K(\cdot)$. The kernel allows for observations to be assigned more or less weight based on their proximity to the cut-off. A kernel is a function which integrates to 1, and which defines how much weight to assign to observations at specific points of the density. A valid kernel then must integrate to 1 and be non-negative. Often, but not always, kernels are symmetric. Commonly used kernels are described in Figure 3.6. In RD designs a triangular kernel is often used, which gives the largest weight to observations which are closest to the cut-off, and which then declines linearly away from this point. Cattaneo and Titiunik (2022) suggest that triangular kernels have MSE optimal properties, while uniform kernels, which give identical weights to each observation local to the cut-off are also often used as these minimize the variance of local polynomial estimators, resulting in narrower confidence intervals.

Figure 3.6: Common Kernel Densities



Selection of Bandwidth The selection of the bandwidth in regression discontinuity estimators is an area with significant research advances in the last decade. The selection of bandwidth h implies that estimation will proceed using *only* observations who fall within the range $x_i \in [\kappa - h, \kappa + h]$. The primary reason for using a larger-than-infinitesimal bandwidth is, of course, sample size. This is a perfect example of the bias-variance trade-offs we sometimes come across in econometrics. While we would like to use only those observations who are *just* above of below the cut-off, if we restrict to too small a sample, estimates will be too imprecise to permit any constructive inference.

Fortunately, there is a considerable amount of work on how to optimally balance this trade-off. Early work by Imbens and Kalyanaraman (2012), sometimes called first generation bandwidth estimators, provide specific guidelines for bandwidth choice.[5] The plug-in estimator for h provides a formula to determine the optimal bandwidth based on, among other things, the sample size available. This formula explicitly recognises the bias-variance trade-off discussed above, depending (negatively) on the bias and (positively) on the variance. The suggested formula for h proposed by Imbens and Kalyanaraman (2012) is:

$$\hat{h}_{IK} = \left(\frac{\widehat{V}_{IK}}{2(p+1)\widehat{B}^2_{IK} + \widehat{R}_{IK}} \right)^{\frac{1}{(2p+3)}} \times n^{\frac{-1}{(2p+3)}}, \tag{3.30}$$

where n is the sample size, p is the degree of the polynomial included on each side of the discontinuity, $\widehat{V}$ is an estimate of the variance of the RD parameter $\hat{\tau}$, $\widehat{B}$ is an estimate of the bias of this parameter, and $\widehat{R}$ is a regularisation term to avoid small denominators when the sample size is not large. Alternatively, Imbens and Kalyanaraman (2012) discuss a manner of calculating optimal h using a cross-validation technique which determines the optimal bandwidth based on the particular sample size of an empirical application (additional details and an example can also be found in Ludwig and Miller (2000)).

The bandwidth $\hat{h}_{IK}$ will lead to an MSE optimal estimator for the parameter τ, but this relies on the underlying estimates for the variance, $\widehat{V}_{IK}$, the bias, $\widehat{B}_{IK}$ and the regularization term. While Imbens and Kalyanaraman (2012) propose estimates for these quantities, the estimates themselves rely on an initial bandwidth, which is itself not optimally chosen. This was followed up by more recent work (Calonico et al., 2014a) which has provided enhancements to the plug-in bandwidth of (Imbens and Kalyanaraman, 2012), suggesting

$$\hat{h}_{CCT} = \left(\frac{\widehat{V}_{CCT}}{2(p+1)\widehat{B}^2_{CCT} + \widehat{R}_{CCT}} \right)^{\frac{1}{(2p+3)}} \times n^{\frac{-1}{(2p+3)}}, \tag{3.31}$$

where $\widehat{V}_{CCT}$, $\widehat{B}_{CCT}$, and $\widehat{R}_{CCT}$ are consistent estimates of their population counterparts, while also using MSE-optimal bandwidths in the generation of these estimates. The precise formulae for these estimates are provided in the appendix of Calonico et al. (2014a) though are quite cumbresome. Fortunately, all of these optimal bandwidth algorithms are available in statistical programming languages such as Stata and R (see for example Calonico et al. (2014b)) so the stability of estimates to different techniques can be examined quite simply. A website is maintained by the authors of this and other related papers at providing a huge amount of useful related econometric material and information about computational implementations at `https://rdpackages.github.io/`.

[5] Packages to implement this are available in Stata and SAS to select the optimal bandwidth. Similar programs also exist for R, MATLAB, and most other computer languages in which econometric estimators are run. More recent optimal bandwidht choice packages described below are provided by the original authors for R, Stata and Python.

Bringing The Ingredients Together With all of the preceding ingredients in hand – a polynomial degree, a kernel and an optimal bandwidth, estimation of the treatment effect in RDDs consists of comparing conditional expectations at the limits on either side of the cut-point. On the left-hand side of the cut-off,

$$\hat{\boldsymbol{\beta}}_{-} = \arg\min_{\beta} \sum_{i=1}^{N} \mathbb{1}\{X_i < \kappa\} \left[y_i - \beta_0 - \beta_1(X_i - \kappa)\right]^2 K\left(\frac{X_i - \kappa}{h}\right)$$

where $\hat{\boldsymbol{\beta}}_{-}$ refers to the vector of parameter estimates $\hat{\beta}_{+,0}, \hat{\beta}_{+,1}$ on the left-hand side of the cut-off, and here we are using a linear polynomial, $p = 1$. Similarly, on the right-hand side of the cut-off:

$$\hat{\boldsymbol{\beta}}_{+} = \arg\min_{\beta} \sum_{i=1}^{N} \mathbb{1}\{X_i \geq \kappa\} \left[y_i - \beta_0 - \beta_1(X_i - \kappa)\right]^2 K\left(\frac{X_i - \kappa}{h}\right).$$

Based on these estimates, the regression discontinuity estimate $\hat{\tau}$ the difference of the intercept at the cut-off point:

$$\hat{\tau} = \hat{\beta}_{+,0} - \hat{\beta}_{-,0}. \tag{3.32}$$

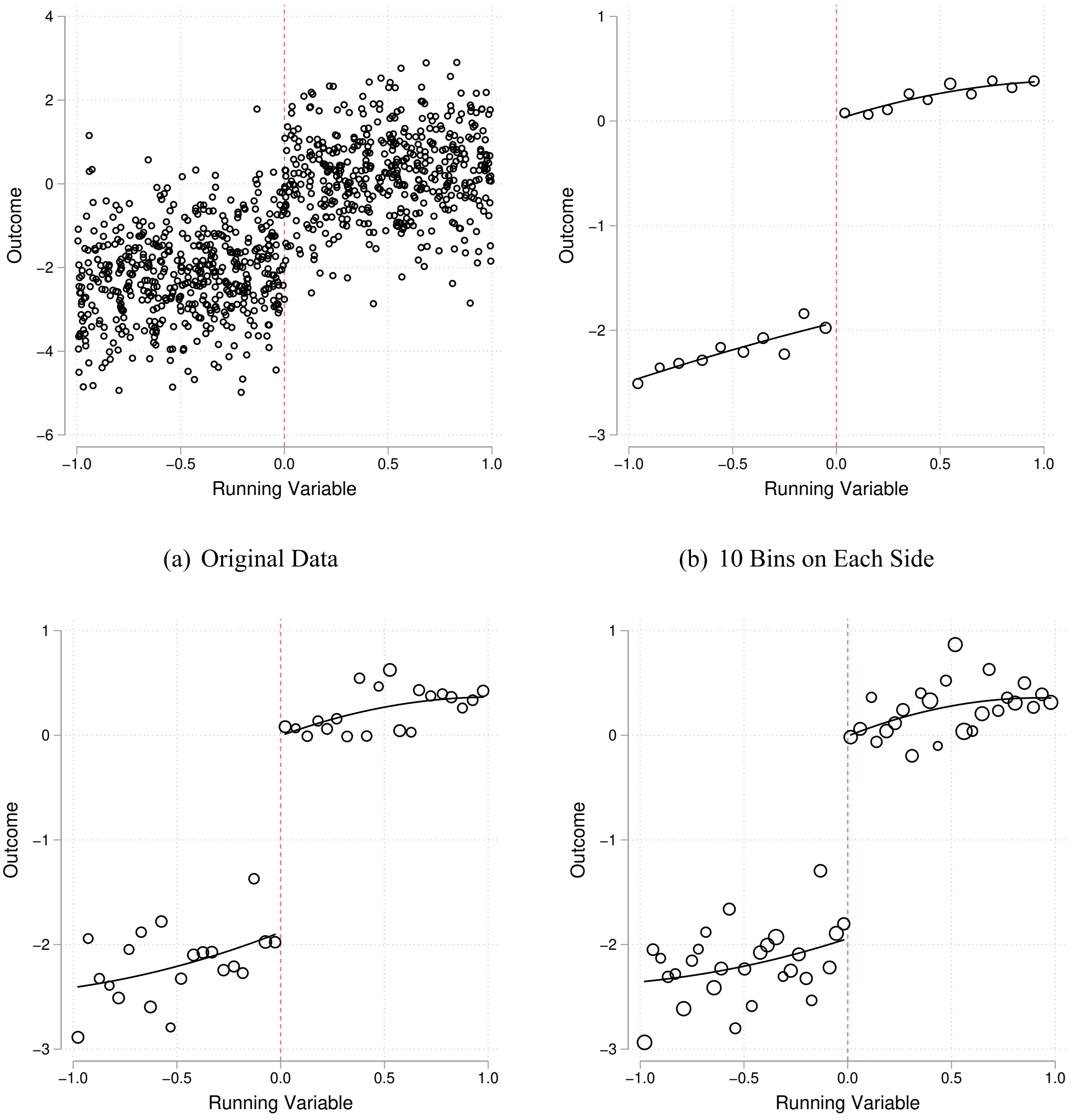
What is nice about this estimate is that it is clear that we are interested in the intercepts on either side of the cut-off, ie the regression estimates at the points where the discontinuity occurs. Under the assumption of local unconfoundedness, $\hat{\tau}$ from equation 3.32 is a consistent estimate for τ from equation 3.26.

The theory behind the estimation of the confidence intervals for this parameter $\hat{\tau}$ is not trivial. As we are interested in estimates directly at the threshold of the regression discontinuity, there is bias due to smoothing of the regression approximation at this point. There is a recent and large body of work on the estimation of standard errors and confidence intervals, with recent implementations often favouring the "robust-bias corrected" confidence intervals proposed in Calonico et al. (2014a). Additional detail can be found in the overview paper of Cattaneo and Titiunik (2022, section 3.2).

Graphical Representations

Arguably, one of the reasons why RD is a successful identification strategy is that it often leads to visually striking representations of the causal effect under study. A shift owing to some arbitrary crossing rule in the running variable generates exogenous variation in the exposure to some particular phenomenon, and this shift can be graphed quite simply in two dimensions. When examining how some outcome of interest moves along the support of the running variable, we can observe both a general pattern describing the relationship between the running variable and the outcome of interest, and importantly, visually inspect for any discontinuities at the assignment thresold.

Figure 3.7: Graphical Representation in an RD Design



(a) Original Data

(b) 10 Bins on Each Side

(c) 20 Bins on Each Side

(d) Optimal Bin Selection

Thus, generally, studies involving a RDD will plot the underlying variation between the running variable and outcome of interest (as well as potentially other patterns described later in this section). Generally, rather than plotting the full dataset, some smoothed function is plotted of averages of outcomes at various points of the running variable. For example, consider the simulated data presented in Figure 3.7. Here a simple discontinuity[6] exists when the running

[6] This is simulated as:

$$y_i = -2 + 0.5w_i + 0.03w_i^2 + 2Treat_i + \varepsilon_i$$

where w_i is a uniform variable with support on [-2,2], $\varepsilon_i \sim \mathcal{N}(0,1)$ and $Treat_i \equiv \mathbb{1}\{w_i > 0\}$, for $i \in \{1,\ldots,1000\}$.

variable plotted on the horizontal access moves from negative to 0 or above. In panel (a) the full data is plotted, where the discontinuity is clearly visible, though the nature of the jump is somewhat disguised by the underlying variation at each point. Remaining panels (b)-(d) present alternative plots, commonly referred to as “RD plots”, where instead of presenting raw data, binned averages are presented in varying numbers of points. In these graphs, each point refers to the average outcomes in small ranges of the running variable, containing mutually exclusive groups of individuals. Here, the size of points refers to the number of individuals contained in each group, and on top of each scatter plot, a quadratic fit of the averaged data is plotted.

Frequently, these bins are chosen arbitrarily (for example using 10 and 20 bins as in panels (b) and (c) of Figure 3.7). However, there are optimal ways to determine bins and generate RD plots. The work of Calonico et al. (2015) provides a data-driven rule for bin selection (as well as allowing for the suggestion of an optimally defined polynomial fit in the graph), which, in the case of bin estimates, are chosen to be evenly spaced or quantile spaced. Evenly spaced refers to bins that are spaced at an equal absolute difference in the running variable, while quantile spaced refers to bins which are spaced such that they are distributed evenly across percentiles of the observed data (taking into account that data may be more sparse at different values of the running variable). These bins are optimally chosen in a way which minimizes the mean-squared error of the regression function describing the relationship between binned averages and the outcome variable of interest. The mean squared error measures the distance between observed averages and the regression-based prediction, and can be shown to be a sum of the variance of the estimate and the square of the bias. Thus, these optimal bins act to trade-off lower variance and lower bias. Full details of these optimal bins, as well as optimal selection of polynomial fits, can be found in Calonico et al. (2015), and computational implementations of this procedure are widely available.

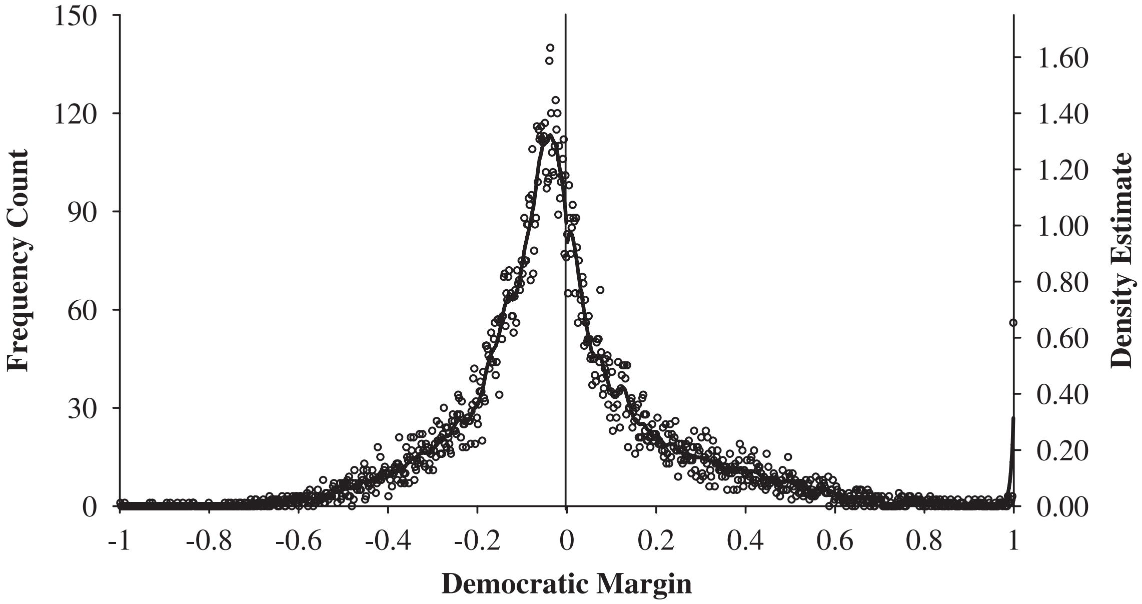
These binned RD plots are generally presented to document the underlying relationship between the running variable and the outcome of interest. However, these plots can also be documented for variables which one would *expect* to be balanced around the RD cut-off. For example, if some measure is available for baseline outcomes of individuals capturing characteristics prior to their assignment to the RD treatment, graphs can be generated to examine whether any discontinuities in baseline outcomes are observed at the point of the discontinuity. If the assumption of local unconfoundedness really is met, one would not expect a similar jump in baseline outcomes at this point. These graphs can thus be used to visually assess whether the assumption of local unconfoundedness is reasonable. We turn to this point more formally below.

3.2.4 Assessing Unconfoundedness

The continuity argument that we used to show that the RD approach estimates a treatment effect suggests a way of testing the underlying assumption. If variation in x around the discontinuity is "as good as" random, then it should also be the case that other variables do not jump at this discontinuity. This is analogous to a balance or placebo test often implemented prior to analyzing data from a randomized, controlled trial (Imbens and Wooldridge, 2009).

A simple way to implement this is to use the same specification as in the outcomes equation, but use instead as a dependent variable some "exogenous" covariate Z_i and test $\lim_{x\downarrow\kappa} E(z_i|x > \kappa) - \lim_{x\uparrow\kappa} E(z_i|x < \kappa) = 0$. If a discontinuity is found in a covariate z_i, this provides evidence that the assumptions underlying the RD design do not hold, even if it is in principle possible to address this by controlling for the covariate in question. For example, Urquiola and Verhoogen (2009) study a RD design which uses class size caps to estimate the effect of class size on children achievement in Chile. They show that in this context parental education and income drop discontinuously at the cutoff, which suggests that better educated parents choose schools where classes are smaller.

Figure 3.8: McCrary test of heaping of running variable (vote shares)



Another tests suggested by McCrary (2008) consists in estimating non parametrically the density of the forcing variable (e.g. through kernel regression) and testing whether it presents some discontinuity around the threshold, i.e. whether $\lim_{x\downarrow\kappa} f_X(x) - \lim_{x\uparrow\kappa} f_X(x) = 0$. If a discontinuity is found in the density of x, then it is likely that individuals were able to manipulate precisely x to choose on which side of the cut-off they were located (e.g. income around "jumps" in the marginal tax rate Kleven and Waseem (2013)). This would cast serious doubt on the RD strategy. Figure 3.8 displays the logic of the test. If there were manipulation of the running variable (in this example, vote share) we may expect to see a heaping of election winners with vote shares just above 50%. This would be evidence in favour of vote buying or some other ballot manipulation, and strong evidence *against* the validity of a local unconfoundedness assumptions. In practice, we see little statistical evidence to suggest that such heaping occurs in this example.

3.2.5 Regression Kink Designs

The regression discontinuity design discussed in previous sections is based on the idea that an external effect creates a discontinuous jump in the likelihood of receiving treatment at a particular point. Another set of methodologies exist when, rather than an appreciable jump in *levels*, we may expect an appeciable change in the *slope* of a relationship at a particular point. These "regression kink designs" are very closely related to the RDDs discussed above, however now we are more interested in the sharp change in the first differential, rather than the level of the variable itself. Examples of kinks from the economic literature include changes in rates of unemployment benefits by time out of work (Landais, 2015), changes in drug reimbursement rates (Simonsen et al., 2016) and various other applications (see table 1 from Ganong and Jäger (2014) for a more exhaustive list).

Card et al. (2015) provide extensive details on the estimation methods and assumptions underlying the regression kink design. Many of the considerations, such as bandwidth calculation and polynomial order are very similar to those in regression discontinuity designs (see also Calonico et al. (2014a) who extend their RDD discussion to the RKD case). In practice, the regression kink design consists of estimating the change in the slope of the outcome variable of interest y_i at the discontinuity:

$$y_i = \beta_0+\beta_1^+ D_i(x_i-\kappa)+\beta_1^-(1-D_i)(x_i-\kappa)+\beta_2^+ D_i(x_i-\kappa)^2+\beta_2^-(1-D_i)(x_i-\kappa)^2+\varepsilon_i \quad (3.33)$$

where here D_i is a binary variable taking 1 when located to the right of the kink, and zero otherwise. Here we are assuming a quadratic functional form, but again, this is can be generalised to other polynomial orders.[7] In order to calculate the treatment effect of the change in exposure, we calculate the RKD estimator as:

$$\hat{\tau}_{RKD} = \frac{\widehat{\beta}_1^+ - \widehat{\beta}_1^-}{\widehat{\gamma}_1^+ - \widehat{\gamma}_1^-}$$

[7] A useful discussion of how to optimally choose polynomial orders is available in Card et al. (2015), who also provide a pointer to other results.

where the estimates of γ are generated by running a similar regression as in equation 3.33, however replacing the outcome variable y_i with the treatment variable. These coefficients capture the corresponding change in the slope of the treatment variable at the discontinuity point. In many cases, the values in the denominator may be known constants, if, for example, they are based on explicit marginal rules, and in these cases rather than estimates, the actual values should be used.

The regression kink set-up relies on similar types of assumptions as those in a regression discontinuity. Namely, we require that no other variables of relevance change their slope at the kink point, and there should be no manipulation of the running variable around the kink point suggestive of people strategically sorting in to points to be eligible for benefits on either side of the cut-off. Fortunately, as is the case with RDDs, these assumptions can be probed with some of the methods described in the previous sub-section.

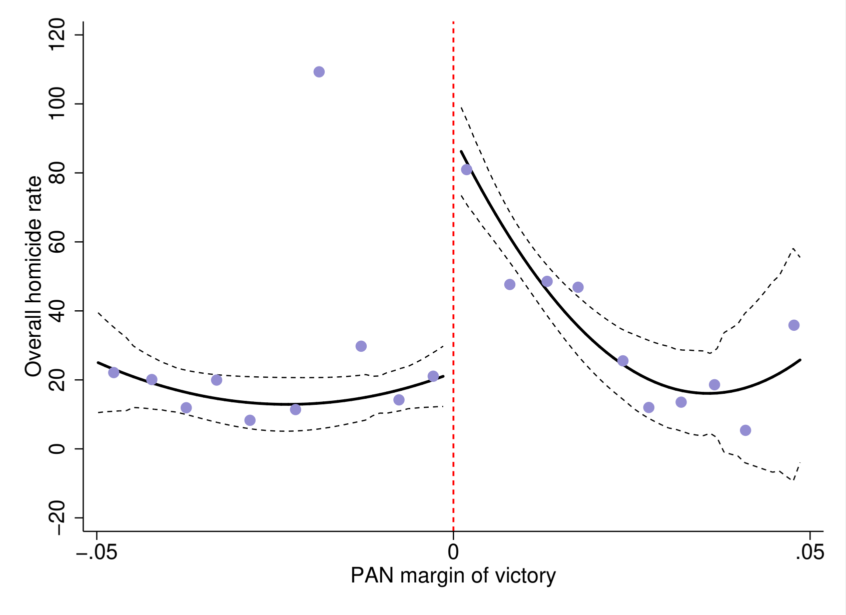
Empirical Exercise 3: Trafficking Networks and the Mexican Drug War

This exercise will have two parts. An applied part, and a part we will simulate ourselves.

The first part of the class (question A) will look at the paper "Trafficking Networks and the Mexican Drug War", by Dell (2015). Her paper examines the effect of Mexican anti-drug policy on drug related violence. She exploits variation in the mayor's party following elections, and uses close elections to estimate using a regression discontinuity design. The PAN party has implemented a number of large-scale anti-trafficking measures, and she examines whether these policies have an effect on drug violence. For further background, the paper is very interesting reading! For part 1, you are provided with the dataset `DrugTrafficking.dta`, which has variables measuring vote share in close elections (only close elections are included), homicides and the rate of homicides, as well as whether the election was won by PAN. A graphical result from the paper (which you will replicate yourselves) is presented below.

For the second part (question B), we will simulate our own data, to examine how regression discontinuity performs when we know the exact data generating process (DGP). Simulation is useful exercise in examining the performance of an estimator in recovering a known parameter: something we only have if *we* have control of the unobservables.

Replication of figure 4 panel B of Dell (2015) "Trafficking Networks and the Mexican Drug War", American Economic Review 105(6):1738-1779



Questions:

(A) Estimating a Regression Discontinuity with Dell (2015) Open the dataset DrugTrafficking.dta, and run the following regression, as per Dell's equation 1:

$$\begin{aligned} HomicideRate_m = \beta_0 + \beta_1 PANwin_m + \beta_2 PANwin_m \times f(VoteDif_m) \\ +\beta_3(1 - PANwin_m) \times f(VoteDif_m) + \varepsilon_m \end{aligned} \tag{3.34}$$

$PANwin_m$ is a binary term for whether PAN won in the close election, while the interaction terms are functions of vote shares on either side of the close election margin, allowing for this "running variable" to behave differently on each side of the discontinuity. In each case we will use the variable $HomicideRate_m$, the rate of homicides at the level of the municipality, as our outcome variable of interest.

1. Run the regression using a linear function for $f(VoteDif_m)$ on each side of the discontinuity.

2. Run the regression using a quadratic function for $f(VoteDif_m)$ on each side of the discontinuity. This will require two terms (linear and squared) on each side of the discontinuity.

3. Replicate the figure on the previous page (panel 4 B from Dell's paper). There is no need to worry about formatting, nor plotting the confidence intervals which are displayed as dotted lines. Note that each point is the average homicide rate in vote share bins of 0.005. You can plot the solid lines on either side of the discontinuity using a quadratic (for example qfit).

4. Why do we focus only on the range of vote margins of -0.05 to +0.05?

(B) Simulating a Regression Discontinuity In this question, we will simulate a discontinuous relationship, and examine how using a local linear regression to capture the discontinuity is appropriate to capture the true effect when the relationship between the running variable (x) and the outcome variable (y) is not linear. We will refer to figure 5 in the notes to simulate our data. This is based on the following DGP:

$$y = 0.6x^3 + 5w + \varepsilon$$

Here y is the outcome variable, x is the running variable, and w is the treatment variable. Treatment will only be received by individuals for whom $x \geq 0$, so w is defined as equal to 1 if $x \geq 0$ and 0 if $x < 0$.

1. Simulate 100 data points which follow the above specification. Note that for this specification, both x and ε are assumed to be drawn from a normal distribution, with mean 0 and standard deviation 1. In Stata, these can be generated the `rnormal()` function, for example, `gen epsilon = rnormal()`. The `set obs` command can be used to define the number of observations to be simulated.

2. Replicate figure 5 from the notes. Do not worry about style. If you want your pseudorandom numbers to exactly replicate those from the notes, before drawing the numbers, use the command `set seed 110`.

3. Estimate the coefficient on the treatment effect w using a linear control for the running variable while concentrating on the observations in the range $x \in (-2, 2), x \in (-1.9, 1.9), \ldots, x \in (-0.1, 0.1)$. Estimation of the effect should use a regression following the above function for y. You can capture the running variable using the same linear trend on both sides, so only need to let x enter the regression linearly, and with no interaction term. This will result in 20 different estimates (one for each set of x ranges). Feel free to display these as you wish, though a graph may be useful in visualising them easily.

   Hint: Rather than doing this all by hand, it may be useful to use a loop! As an example, consider running a regression of y on x only for those observation who have x greater than a series of numbers, and saving the coefficient on x from each regression as a seperate observation in the variable `coefficients`, and the x cutoff from each regression in the variable `cutoff`:

   ```
   gen coefficients = .
   gen cutoff = .
   local i = 1
   foreach num of numlist 0.1(0.1)2 {
       reg y x if x > `num'
       replace coefficients = _b[x] in `i'
       replace cutoff = `num' in `i'
       local i = `i'+1
   }
   ```

   You will need to apply this code to the specific example in question 3, which will require some modifications!

4. What do the above results tell you about the performance of RDD using local linear regressions? Is there some theoretical guidance on how to determine the optimal bandwidth? If so, what are the considerations in making this choice?

Chapter 4

Testing, Testing: Hypothesis Testing in Quasi-Experimental Designs

Required Readings
Romano et al. (2010) (section 8 only)

Suggested Readings
Anderson (2008)
Dobbie and Fryer (2015)
Gertler et al. (2014)

The nature of frequentist stastical tests implies that we will at times make mistakes. Indeed, this is built directly into the framework which we have also used in inference up to this point. When we refer to a parameter being significant *at 95%*, we mean that if we were able to repeat this test many times, in 5% of those we would incorrectly reject the null hypothesis. In general, this is not a problem as long as our inference respects the nature of these tests, and our findings are taken in light of this chance. However, in this final section of the course we will consider a number of situations in which this *may* be a problem. The first: how to consider hypothesis tests when we have multiple dependent variables is a technical issue for which, fortunately, there are many solutions. The second, abuse of the notion of frequentist testing owing to incentives to report a significant result is a deeper problem related to research in social sciences, on which a lot of attention is only recently being placed.

If researchers are selectively more likely to report positive results, or if there are strong incentives in place which mean that statistically significant findings are more valuable, the nature of our traditional hypothesis tests breaks down. At its most extreme, the crux of this problem is summed up precisely by Olken (2015). As he states:

> "Imagine a nefarious researcher in economics who is only interested in finding a statistically significant result of an experiment. The researcher has 100 different variables he could examine, and the truth is that the experiment has no impact. By construction, the researcher should find an average of five of these variables statistically significantly different between the treatment group and the control group at the 5 percent level—after all, the exact definition of 5 percent significance implies that there will be a 5 percent false rejection rate of the null hypothesis that there is no difference between the groups. The nefarious researcher, who is interested only in showing that this experiment has an effect, chooses to report only the results on the five variables that pass the statistically significant threshold."

Olken (2015), p. 61.

And indeed, this problem is certainly not new, and is not isolated to only the social sciences! A particularly elegant (graphical) representation of a similar problem is described in the figure overleaf.

In this section we will, briefly, recap the ideas behind the basic hypothesis test and the types of errors and uncertainty that exists. Then we will discuss how these tests can be extended to take into account various challenges, including very large sample sizes, and the use of multiple dependent variables. We will then close discussing one particular way which is increasingly used to avoid concerns about the selective reporting problem described above, namely, the use of a pre-analysis plan to pre-register analyses before data are in hand, thus removing so called "researcher degrees of freedom" from analysis.[1]

4.1 Size and Power of a Test

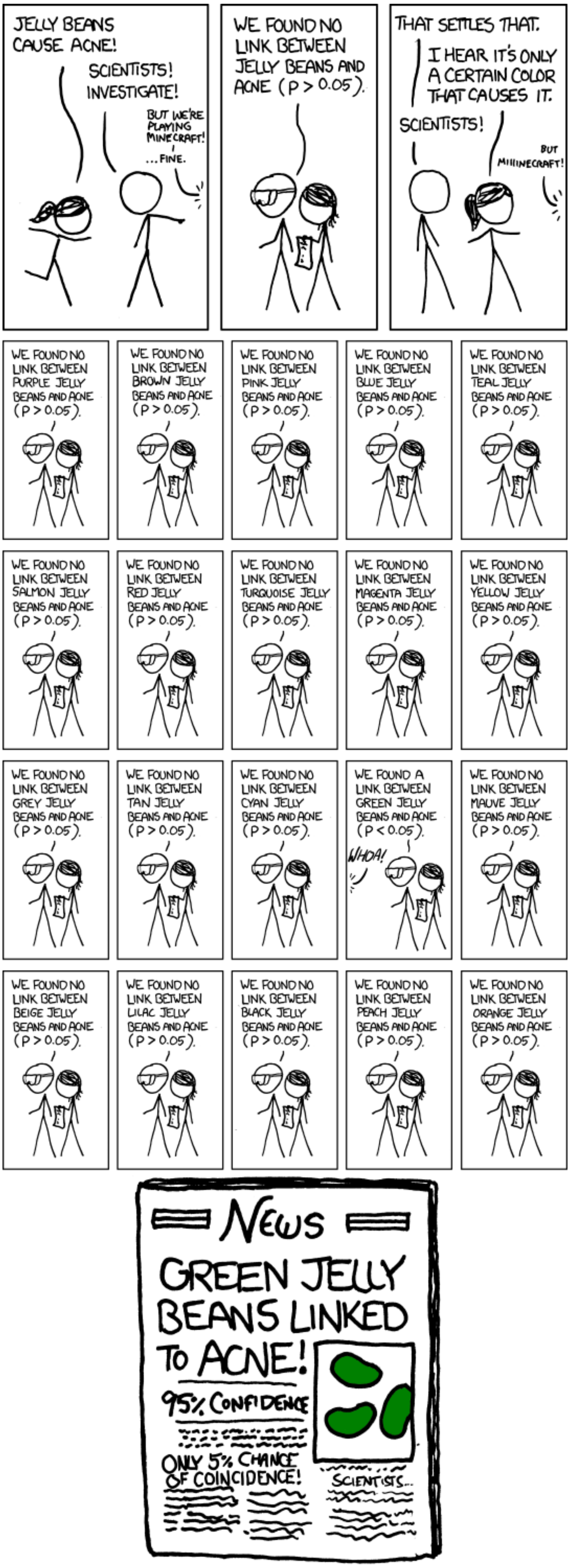
In order to think about hypothesis testing and the way that we would like to be able to classify treatment effects, we will start by briefly returning to the typical error rates from simple hypothesis tests. Let's consider a hypothesis test of the type:

$$H_0 : \beta_1 = k \qquad \text{versus} \qquad H_1 : \beta_1 \neq k.$$

In the above, our parameter of interest is β_1, and k is just some value which we (the hypothesis tester) fix based on our hypothesis of interest.

[1] For some interesting additional discussion on these issues refer to work by Andrew Gelman and colleagues (for example Gelman and Loken (2013)). Andrew Gelman also has a blog where he provides frequent interesting analysis on issues of this type (`http://andrewgelman.com`).

Figure 4.1: A Funny Comic but a Serious Problem (Munroe, 2010)


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

Given that β_1 is a population parameter, we will never know with certainty if the equality in H_0 (the "null hypothesis") holds. The best that we can do is ask how likely or unlikely is it that this hypothesis is true given the information which we have available to us in our sample of data. In simple terms, producing an estimate for β_1 which is very far away from k will (all else constant) give us more evidence to believe that the hypothesis should not be accepted.

Classical hypothesis testing then consists of deciding to reject or not reject the null hypothesis given the information available to us. Although we will never know if we have correctly or incorrectly rejected a null, there are four possible states of the world once a hypothesis test has been conducted: correctly reject the null; incorrectly reject the null; correctly fail to reject the null; incorrectly fail to reject the null. Two of these outcomes (the underlined outcomes) are errors. In an ideal world, we would like to perfectly classify hypotheses, never committing either types of the errors above. However, given that in applied econometrics we never know the true parameter β_1, and that hypothesis tests are based on stochastic (noisy) realizations of data, we can never simultaneously eliminate both types errors.

4.1.1 The Size of a Test

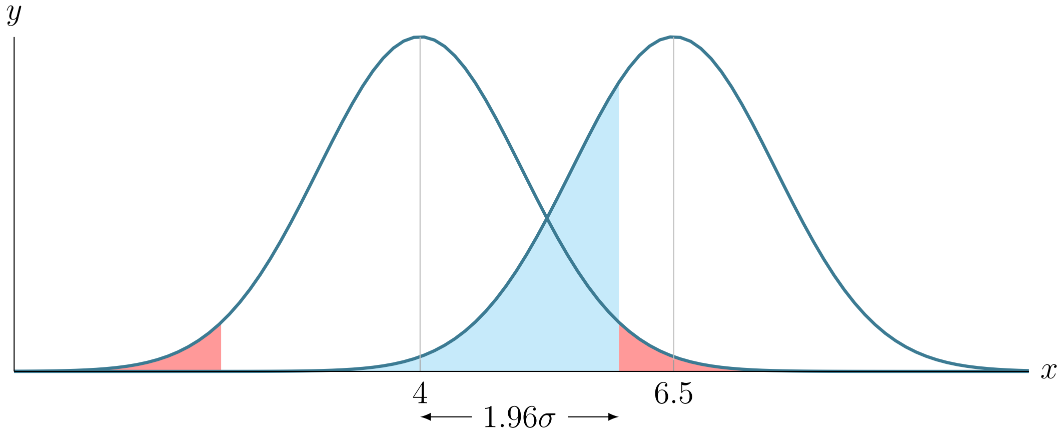
The size of a test refers to the probability of committing a Type I error. A type I error occurs when the null hypothesis is rejected, even though it is true. In the above example, this is tantamount to concluding that $\beta_1 \neq k$ despite the fact that β_1 actually *is* equal to k. Such a situation could occur, for example, if by chance a sample of the population is chosen who all have higher than average values of β_1

The rate of type I error (or the size of the test) is typically denoted by α. We then refer to $1 - \alpha$ as the confidence interval. Typically we focus on values of α such as 0.05[2], implying that if we repeated a hypothesis 100 times (with different samples of data of course) then in 5 out of every 100 times we would incorrectly reject the null if the hypothesis were actually true. In cases where we run a regression and examine whether a particular parameter is equal to zero, setting the size of the test equal to 5 implies that in 5% of repeated tests we would find a significant effect even when there is no effect.

In figure 4.2, the red regions of the left-hand curve refer to the type I error. Assuming that the true parameter β_1 is equal to 4 and the distribution of the estimator for the parameter $\widehat{\beta}_1$ is normal around its mean, we will consider as evidence against the null any value of $\widehat{\beta}_1$ which is outside of the range $4 \pm 1.96\sigma$ (where σ refers to the standard deviation of the distribution of the estimator). We do this knowing full well that in certain samples from the true population

[2] Lehmann and Romano (2005, p. 57) report that standard values for α were originally chosen given that it allowed for fewer statistical tables to be produced when critical values where generally tabulated. Currently, the ease of generating critical values with a computer is so easy that this is no longer necessary, but the practice has stuck.

Figure 4.2: Type I and Type II Errors



(in 5% of them to be exact!) we will be unlucky enough to reject the null *even though* the true parameter is actually 4. Of course, there is nothing which requires us to set the size of the test at $\alpha = 0.05$. If we are concerned that we will commit too many type I errors, then we can simply increase the size of our test to, say, $\alpha = 0.01$, effectively demanding stronger evidence from our sample before we are willing to reject the null.

4.1.2 The Power of a Test

Power in Detecting Impacts Versus a Scalar Null Hypothesis

These discussions of the size of a test and type I errors are entirely concerned with incorrectly rejecting the null when it is true. However, they are completely silent on the reverse case: failing to reject the null when it is actually false. This type of error is referred to as a type II error. We define the power of a statistical test as the probability that the test will correctly lead to the rejection of a false null hypothesis. We can then think of the power of a test as the ability that a test has to detect an effect if the effect actually exists. For example, in the above example imagine if the true population parameter were 4.01. It seems unlikely that we would be able to reject a null that $\beta_1 = 4$, even though it is not true. As we will see below, considerations of the power of a test are particularly frequent when deciding on the sample size of an experiment or RCT with the ability to determine a minimum effect size.

The statistical power of a test is denoted by $1 - \beta$, where β refers to the Type II error. Often, you may read that tests with power of greater than 0.8 (or $\beta \leq 0.2$) are considered to be powerful. An illustration of the concept of statistical power is provided in figure 4.2. Imagine that we would like to test the null that $\beta_1 = 4$, and would like to know what the power of the test would be if the actual effect was 6.5. This amounts to asking, over what portion of the distribution of the true effect (with mean 6.5), will the estimate lie in a zone which causes us not to reject the null that $\beta_1 = 4$. As we see in figure 4.2, there is a reasonable portion of the distribution (the shaded blue portion) where we would (incorrectly) not reject the null that $\beta_1 = 4$ if the true effect were equal to 6.5.

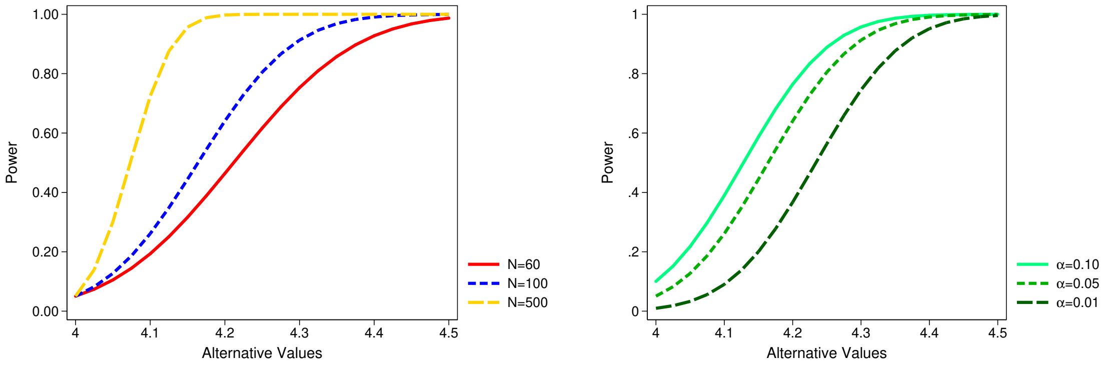
In looking at figure 4.2, we can distinguish a number of features of the power of a test. Firstly, the power of a test will increase as the distance between the null and the true parameter increase. This is to say that we would have greater power when considering 7 to $\beta_1 = 4$ than 6.5 to $\beta_1 = 4$ (all else equal). Secondly, we will have greater power when the standard error of the estimate is smaller. As the standard error gives the dispersion of the two distributions, as these dispersions shrink, we will be more able to pick up differences between parameters. As the standard error depends (positively) on the standard deviation of the estimate and (negatively) on the sample size, the most common way to increase power is by increasing the sample size. Finally, we can see that by *increasing* the size of the test (ie changing the significance level from $p = 0.05$ to $p = 0.10$), that this *increases* the power of the test. We can see this in figure 4.2, as by increasing the red area (that is, increasing the likelihood of making a type II error), we shrink the size of the blue area (we reduce the likelihood of a type I error). Here we see an interesting and important fact: we can not simultaneously both increase the power and reduce the size of the test simply by changing the significance level. Indeed, the opposite is true, as there exists a trade-off between type I and type II errors in this case.

These three facts can be summed up in what we know as a “power function”. Although figure 4.2 only considers one value (6.5), we can consider a similar power calculation for a whole range of values. The power function summarises for us the power of a test given a particular true value, conditional on the sample size, standard deviation, and value for α. In particular, imagine that we have a parameter β_1 which we believe follows a t-distribution, and for which we want to test the null hypothesis that $H_0 : \beta_1 = 4$. Let’s imagine now that the alternative is actually true, and $\beta_1^T = \theta$, where we use β_1^T to indicate it is the true value. We can thus derive the power at $\alpha = 0.05$ using the below formula, where we use the critical value of 1.64 from the t-distribution:

$$\begin{aligned} B(\theta) &= Pr(t_{\beta_1} > 1.64 | \beta_1^T = \theta) \\ &= Pr\left(\frac{\hat{\beta}_1 - 4}{\sigma^2/\sqrt{N}} > 1.64 \middle| \beta_1^T = \theta\right) \\ &\approx 1 - \Phi\left(1.64 - \frac{\theta}{\sigma^2/\sqrt{N}}\right). \end{aligned} \tag{4.1}$$

where the final line comes from using the normal distribution as an approximation for the t-distribution when N is large. The idea of this forumla is summarised below in the power functions described in figure 4.3. In the left-hand panel we observe the power function under varying sample sizes (and values for θ), and in the right-hand panel observe the power functions where the size of the test changes (and once again, for a range of values for θ).

Figure 4.3: Power Curves



(a) Varying Sample Size

(b) Varying Significance Level

Power in Detecting Differences Between Groups

A particular case where power is often discussed is in the design of RCTs, where one, ex-ante, must decide on a required sample size, with larger sample sizes implying larger costs in treatment, enumeration, and so forth. Generally, the sample size is chosen to ensure a power to detect some minimum desired effect size between a treatment and a control group. A nice discussion of this can be found in Athey and Imbens (2017, pp. 102–104). If you are ever in the situation of implementing an RCT, it is worth reading this carefully, along with the references cited therein such as Cohen (1988); Murphy et al. (2014) or other readings discussed in Chapter 1 of these notes. We will briefly review this consideration below, following the notation of Athey and Imbens (2017).

Let τ signify the true treatment effect of receiving some treatment, and assume that γ refers to the proportion of individuals receiving treatment, with the remaining proportion $1-\gamma$ acting as control units. For simplicity, assume that the variance of outcomes is the same, indicate by σ^2. In what follows, subscript t will refer to units receiving treatment, and subscript c will refer to units acting as controls. Generally, when conducting power calculations, we wish to determine the minimum sample size necessary, $N = N_c + N_t$, to assure a rejection probability of at least β given that the alternative hypothesis is true, and the true treatment effect is τ. We can start from the standard result that the difference in means between treatment and control minus the true treatment effect divided by the standard error of this difference is approximately a standard normal distribution:

$$\frac{\bar{Y}_t - \bar{Y}_c - \tau}{\sqrt{\sigma^2/N_t + \sigma^2/N_c}} \approx \mathcal{N}(0,1). \tag{4.2}$$

Now, consider the t-statistic which will be tested when examining a null of a zero effect:

$$t = \frac{\bar{Y}_t - \bar{Y}_c}{\sqrt{\sigma^2/N_t + \sigma^2/N_c}}. \tag{4.3}$$

Rearranging the result from 4.2, implies that 4.3 has an approximately normal distribution as:

$$t \approx \mathcal{N}\left(\frac{\tau}{\sqrt{\sigma^2/N_t + \sigma^2/N_c}}, 1\right).$$

Considering the properties of the normal distribution, this implies that the probability of rejecting the null of equality between groups at a signficance level α if the true effect is τ is:

$$\begin{aligned} Pr(|t| > \Phi^{-1}(1-\alpha/2)) \;\; \approx \;\; & \Phi\left(-\Phi^{-1}(1-\alpha/2) + \frac{\tau}{\sqrt{\sigma^2/N_t + \sigma^2/N_c}}\right) \\ & + \Phi\left(-\Phi^{-1}(1-\alpha/2) - \frac{\tau}{\sqrt{\sigma^2/N_t + \sigma^2/N_c}}\right), \end{aligned} \tag{4.4}$$

where Φ refers to the standard normal CDF, and Φ^{-1} its inverse. The second term here is small (strictly smaller than $\Phi(-\Phi^{-1}(1-\alpha/2))$, for example strictly smaller than 0.025 when $\alpha = 0.05$), and as such we will ignore it in what follows.

We wish to ensure a minimum power of $1-\beta$, so from 4.4:

$$1 - \beta = \Phi\left(-\Phi^{-1}(1-\alpha/2) + \frac{\tau}{\sqrt{\sigma^2/N_t + \sigma^2/N_c}}\right)$$

which can be simplified as:

$$\Phi^{-1}(1-\beta) = -\Phi^{-1}(1-\alpha/2) + \frac{\tau\sqrt{N}\sqrt{\gamma(1-\gamma)}}{\sigma},$$

finally allowing us to arrive to a formula for the effective sample size required to detect a minimum treatent effect of τ, depending on the desired α, β, γ and standard deviation σ as:

$$N = \frac{\left(\Phi^{-1}(1-\beta) + \Phi^{-1}(1-\alpha/2)\right)^2}{(\tau^2/\sigma^2)\cdot\gamma\cdot(1-\gamma)}.$$

A brief applied example illustrating such a calculation is provided by Athey and Imbens (2017, p. 104).

4.2 Hypothesis Testing with Large Sample Sizes

While in typical experimental analyses we are much more likely to be concerned about a sample size which is too small to permit precise inference, we should—briefly at least—discuss the reverse case. In some circumstances we will be working with very large samples of data. This is particularly so when using quasi-experimental methods, and for example, administrative datasets. In these cases it may not be at all uncommon to work with millions or even tens of millions of observations.

In these cases, we will likely find that nearly *everything* is significant when conducting hypothesis tests of the sort $\beta = \beta_0$. This is of course not a reflection that the truth surrounding a hypothesis depends on the sample size, but rather a feature of the way we calculate test statistics. As our typical test statistics depend inversely on the standard errors of estimated coefficients, and as these coefficients depend inversely on sample size, then as the sample size grows it is easier for us to find that our test statistic exceeds some fixed critical value.

This fact has been well pointed out and discussed in various important applied texts. Deaton (1997) provides an extremely clear discussion of this phenomena, drawing on a more extensive set of results from Leamer (1978). As the sample size grows, we have increasing quantities of information with which to test our hypotheses. As Deaton (1997) points out, why then should we be content with still rejecting the null hypothesis in 5% of the cases when it is true? As we have seen in the previous section, increasing the sample size increases the power of a test, reducing the likelihood that we commit a type I error. However, as we gain more and more power with the increasing sample size, it seems inefficient to maintain fixed the size of the test, committing equally as many type II errors. Rather, it is suggested by Deaton (1997), Leamer (1978) and others that we should dedicate at least *some* of the additional sample size to reducing the size of the test, lowering the probability of incorrectly rejecting the null. Lehmann and Romano (2005) state this in the sense that when power is very close to 1, α can be reduced without losing very much power at all, suggesting potentially a large gain in size without much cost to power.

In practice, it is suggested that we should set critical values for rejection of the null which increase with the sample size. While the full details of the derivation go beyond what we will look at here[3] the suggestion is actually rather simple. Rather than simply rejecting an F or t test if the test statistic exceeds some critical value, we should reject the test if:

$$F > \left(\frac{N-K}{P}\right)\left(N^{\frac{P}{N}}-1\right) \qquad \text{or} \qquad t > \sqrt{(N-K)\left(N^{\frac{1}{N}}-1\right)},$$

where N refers to the sample size, K the number of parameters in the model, and P the number of restrictions to be tested. Moreover, as Deaton (1997) points out, these values can be approx-

[3] They can be found in Leamer (1978) and are based on Bayesian, rather than classical, hypothesis testing procedures.

imated by $\log N$ and $\sqrt{\log N}$ respectively. Clearly then, these tests set the rejection of the null in a way that it grows with the sample size, and so the rate of type II errors will become increasingly small. For an empirical application in which this methods is employed, see for example Clarke et al. (2016).

4.3 Multiple Hypothesis Testing and Error Rates

In the previous sections we have thought about hypothesis tests where we are interested in conducting a single test, either based on a single parameter (a t-test) or multiple parameters (an F-test). Setting the rejection rate of a simple hypothesis test of this type at α leads to an unequivocal rule with regards to acceptance or rejection of the null, and a similarly clear understanding of the rate of type II errors. exceeds the critical value at α, reject H_0, otherwise do not reject.

However, we may not always have a single hypothesis to test. For example, what happens if we have a single experiment (leading to one exogenous independent variable) which we hypothesise may have an effect on *multiple* outcome variables? This is what we refer to as "multiple hypothesis testing",[4] and it brings about a series of new challenges. To see why, consider the case of a single independent variable and two outcome variables. If we run the regression once using the first outcome variable and test our hypothesis of interest, we will have a type I error rate of α. However, if we then also the regression a second time using the second outcome variable, the chance of making *at least one* type I error in these tests now exceeds α, as both regressions contribute their own risk of falsely rejecting a null. This may have very important consequences for the way that we think about the effect of a policy. If we consider that evidence of an effect of the policy on any variable in a broad class is suggestive that the policy is worthwhile, the accumulation of type I errors will make us more likely to find that a policy is worthwhile as the number of variables examined increases.

More generally, assuming for simplicity that each hypothesis test is independent, the likelihood of falsely rejecting *at least* one null incorrectly in a series of m tests when all the null hypotheses are correct is equal to $1-(1-\alpha)^m$. Thus, if 10 hypotheses relating to 10 outcome variables are tested, the likelihood of at least one true null hypothesis being rejected is $1-(1-0.05)^{10} = 0.401$!

This is clearly problematic, and something that we need to think about. However, before

[4] We should be quite careful in making sure that we understand the difference between a test where we are intersted in knowing if there are various independent variables which may affect a single dependent variable, in which case all we need is an F-test, and one in which a single independent variable may impact various dependent variables. It is the latter which we are concerned with, as in this case we will be estimating various regression models with different outcome variables.

continuing to examine a series of proposed solutions, we will discuss a series of alternative error rates which are relevant when working with multiple hypotheses. When considering multiple, rather than single hypothesis tests, it is not clear that there is only one way to think about the type I error rates associated with hypothesis tests. For example, should we demand that our hypothesis tests with multiple variables should set error rates based on falsely rejecting *any one* of the hypotheses in a group, or the *total percent* of all hypotheses in a family, or some other rejection rate?

This gives rise to different error rates. Among these, the Family Wise Error Rate (FWER), the Generalised FWER (k-FWER), and the False Discovery Rate (FDR). The **Familywise Error Rate (FWER)** gives the probability of rejecting at least one null hypothesis in a family when the null hypothesis is actually true. The **Generalised Familywise Error Rate (k-FWER)** is similar to the familywise error rate, however, now instead of the probability of falsely rejecting at least *one* null hypothesis, it now refers to the probability of rejecting at least k null hypotheses, where k is a positive integer. Finally, the **False Discovery Rate (FDR)** refers to the proportion of all expected "discoveries" (rejected null hypotheses) which are true.

These different error rates are clearly different, with the FWER being more demanding than the FDR. In the family wise error rate, we demand that were we test *all* our multiple hypotheses many times using separate draws from the DGP, only in α% of the cases would we falsely reject any of these hypotheses. On the other hand, with the FDR, we know that with a significantly large number of findings, α% will actually be false. There exist a range of methods to control the FWER or the FDR. The type of method used will depend largely on the context. Where *any* evidence in favour of a hypothesis is instrumental in applied research, it may be most correct to fix the FWER, as this way our error rates take into account the likelihood of falsely rejecting any null. However, although the FWER is more demanding and hence gives rise to stronger evidence where a null is rejected, it should be recognised that there will be circumstances in which the FWER is simply too demanding to work in practice. Mainly, this is the case when the number of hypotheses in a family is so large that it will be very difficult to avoid falsely rejecting any hypothesis. In the sections below we discuss different correction methods to control for these two rates.

4.4 Multiple Hypothesis Testing Correction Methods

4.4.1 Controlling the FWER

There are a number of proposed ways to adjust significance levels or testing procedures to account for multiple hypothesis testing by controlling the FWER. Some of these data from as far back as the early 20th century and are still widely used today. As we will see below, alternative procedures are more or less conservative, with important implications to the power of the test.

In what follows, let's consider a series of S hypothesis tests, which we label $H_1, \ldots, H_S$. Thus, the family of tests consists of S null hypotheses, and we will assume that S_0 of these are true null hypotheses. In the traditional sense, each of the S hypotheses is associated with their own p-value labelled $p_1, \ldots, p_S$.

The earliest type of multiple hypothesis adjustment is the Bonferroni (1935) correction. The Bonferroni correction simply consists of adjusting the rejection level from each of tests in an identical way. Rather than rejecting each test if $p_s < \alpha$, the rejection rule is set to reject the null if $p_s < \frac{\alpha}{S}$. It can be shown that under this procedure, the Family Wise Error Rate is *at most* equal to α (though likely much lower). To see why, consider the following:

$$FWER = Pr\left[\bigcup_{s=1}^{S_0}\left(p_s \leq \frac{\alpha}{S}\right)\right] \leq \sum_{s=1}^{S_0}\left[Pr\left(p_s \leq \frac{\alpha}{S}\right)\right] \leq S_0\frac{\alpha}{S} \leq S\frac{\alpha}{S} = \alpha.$$

In the above, even if all the tested hypotheses are true (ie $S = S_0$) we will never falsely reject a hypothesis in greater than α% of the families of tests.[5] However, this is a particularly demanding correction. Imagine, for example if we are testing $S = 5$ hypotheses, and would like to determine for each whether their exists evidence against the null at a level of $\alpha = 0.05$. In order to do so, we must adjust our significance level, and *only* reject the null at 5% for those hypothesis for which $p_s < 0.01$. It is simple to see that as we add more and more hypotheis to the set of test, the global significance level required to reject each null quickly falls.

However, one benefit of the Bonferroni (1935) correction is that it is extremely easy to implement. It requires no complicated calculations, and can be done 'by eye' even where a paper's authors may not have reported it themselves. Further, this procedure does not require any assumptions about the dependence between the p-values or about the number of true null hypotheses in the family. Of course, this flexibility comes at a cost…We see below how we can increase the efficiency of multiple hypothesis testing by taking these into consideration.

Single-Step and Stepwise Methods

The Bonferroni (1935) correction is an example of a single-step multiple hypothesis testing correction methodology. In these single-step procedures, all hypotheses in the family are compared in one shot a global rejection rate leading to S reject/don't reject decisions. However, there also exists a series of stepwise methods, which rather than comparing all hypotheses at once, begin with the most significant variable, and iteratively compare it to increasingly less

[5] The precise details of the proof of the above rely on Boole's Inequality for the first step. While not necessary for the results discussed in this course, if you would like further details, most statistical texts will provide useful details, for example Casella and Berger (2002).

conservative rejection criterion. The idea of these stepdown methods is that there is an additional chance to reject less significant hypotheses in subsequent steps of the testing procedure (Romano et al., 2010).

One of the most well known of these methods – which similarly maintains the simplicity we observed in the Bonferroni correction – is the Holm (1979) multiple correction procedure. This method begins with a similar idea to the Bonferroni correction, however is less conservative, and hence more powerful (indeed, it is a "universally more powerful" testing procedure, meaning it will reject all the false nulls rejected by Bonferroni, and perhaps more). The idea is that rather than making a one-shot adjustment to α for all S hypotheses, we make a step-wise series of adjustments, each slightly less demanding given that certain hypotheses have already been tested. In the Bonferroni correction then simply consists of rejecting the null for all H_s where $p_s \leq \alpha/S$.

Holm (1979)'s correction proceeds as follows. First, we order the p-values associated with the S hypotheses from smallest to largest:

$$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(S)},$$

and we name the corresponding hypotheses as $H_{(1)}, H_{(2)}, \ldots, H_{(S)}$. We then proceed step-wise, where each of the hypotheses is rejected at the level of α if:

$$p_{(j)} \leq \frac{\alpha}{(S-j+1)} \qquad \forall\, j = 1, \ldots, S. \tag{4.5}$$

Thus, in the limit (for the first test), Holm's procedure is identical to the Bonferroni correction given that the denominator of equation (4.5) equals $S - 1 + 1 = S$. And in the other limit (for the final test), the procedure is identical to a single hypothesis test of size α, given that the denominator of (4.5) is equal to $S - S + 1 = 1$.

Bootstrap Testing Procedures Up to this point in these lectures we have always worked with test-statistics with a closed form solution. For example, a traditional t-test in a regression model is simply calculated using the estimator and its standard error, and both of these have simple analytic solutions (at least when estimating using OLS). However, using an analytical test-statistic with proven desirable qualities is only one possible way to conduct inference. Another, and indeed more flexible, class of inference is based on resampling methods. These methods, which we have alluded to only very briefly when discussing difference-in-differences models, include as a principal component the bootstrap, of Efron (1979). Here we will briefly discuss the idea of a bootstrap estimate for a confidence interval, before showing how we can use a bootstrapped test statistic to produce more efficient multiple hypothesis tests.

The idea of the bootstrap is one of analogy. Normally in hypothesis testing we are interested in the population. However, we only have a single sample from this population, which we assume is representative. The logic behind the bootstrap is to treat the sample as analogous to the true population. Then, by taking many resamples from our original sample, and in each case calculating our parameter of interest, we can build an entire distribution of estimates, giving a range for our point estimate. From the work of Efron (1979) we know that the bootstrap is an asymptotically valid way to approximate the true distribution.

In order to understand a bit more we will introduce some basic notation. Imagine that we have a sample of size N, and parameter of interest we will call β. If we estimate β in the original sample this gives us $\widehat{\beta}$. Now, imagine that we are interested in creating a "new" dataset by taking a re-sample from our original data. This re-sample simply chooses at random N observations from our original dataset *with replacement*. As the sample is taken with replacement (that is to say a single observation from the original sample may be included 0, 1, or multiple times in the re-sample), this leads to a different dataset. Using this new re-sampled dataset we can once again estimate β, leading to a different estimate $\widehat{\beta}^{*1}$. each re-sample is a different dataset. Here we use $*$ to indicate that our estimate comes from a re-sample, and 1 to indicate that it is the first re-sample. Finally, we conduct the above re-sampling procedure (always from the original dataset) $B-1$ more times, resulting in B "new" datasets, and hence B estimates for β, denoted $\widehat{\beta}^{*1}, \widehat{\beta}^{*2}, \ldots, \widehat{\beta}^{*B}$. In order to find the 95% confidence interval for our original estimate $\widehat{\beta}$ we simply order these bootstrap estimates $\widehat{\beta}^{*}$, and find the upper and lower bound using the estimates at quantiles 2.5 and 97.5.

We can also use a bootstrap method to run hypothesis tests and calculate p-values. Imagine, for example, that we wish to calculate the p-value associated with the test that the above parameter $\beta = 0$. Using each of the $b \in B$ bootstrapped estimates we can generate a distribution of t-statistics, where we *impose* that the null is true. Consider the following calculation corresponding to each of the β^{*} terms:

$$t^{*b} = \frac{\widehat{\beta}^{*b} - \overline{\widehat{\beta}^{*}}}{\sigma(\widehat{\beta}^{*})}.$$

Here $\overline{\widehat{\beta}^{*}}$ refers to the average $\widehat{\beta}^{*}$ among all B resamples, and $\sigma(\widehat{\beta}^{*})$ refers to the standard deviation of these estimates. This then results in a distribution of t-statistics using the resampled data which is what we would expect if the true β were equal to zero. All that remains for our hypothesis test then is to compare our actual t-value (from the true estimate $\widehat{\beta}$) with the distribution in which the null is imposed. This actual t-statistic is simply based on our estimate $\widehat{\beta}$, which is standardised using the same standard deviation as above: $t = \widehat{\beta}/\sigma(\widehat{\beta}^{*})$. If the actual t-value, which we will call t, is much higher or much lower than those in the null distribution, we will conclude that it is unlikely that the null hypothesis is true. What's more, we can attach a precise p-value to this hypothesis test. All we need to do is ask "what percent of t-statistics from the null distribution exceed the true t-statistic?" If this proportion is low, it is strong evidence against the null. This results in the following calculation of a p-value, where for simplicity we take the absolute value of the t-statistics given that we are interested in values which are located in either extreme tail of the distribution. We denote this value as p^* to signify that it comes from the bootstrap calculation, and it is reasonably easy to show that $0 \leq p^* \leq 1$, with a lower value of p^* signifying greater evidence against the null. We would typically work with a value such as $\alpha = 0.05$ as a rejection criteria.

$$p^* = \frac{\#\{|t^*| \geq |t|\} + 1}{B + 1}$$

Romano-Wolf Stepdown Testing A final, and particularly efficient, means of fixing the FWER is the Romano-Wolf step-down testing procedure, described in Romano and Wolf (2005a,b). This procedure is increasingly used in the economic literature, for example in Gertler et al. (2014); Dobbie and Fryer (2015). This procedure is based on a bootstrap testing procedure similar to that described above, however correcting for the fact that we are conducting multiple hypotheses at once. It is a step down testing procedure (similar to Holm (1979)), and so considers one hypothesis at a time, starting with the most significant.

Consider the same S hypotheses considered above, ordered again from most to least significant as $H_{(1)}, H_{(2)}, \ldots, H_{(S)}$. For each of these hypotheses we will generate a null distribution of test-statistics using the bootstrap method described above, and B replications. This gives a series of resampling distributions $\boldsymbol{t_1^*}, \boldsymbol{t_2^*}, \ldots, \boldsymbol{t_S^*}$ where each of these is a vector of B values.

The Romano Wolf testing procedure is then based on using the information from all of these re-sampling distributions to correct for the fact that multiple hypotheses are tested at once. For the first hypothesis we construct a new null distribution which, for each of the B resamples takes the *maximum* t-value associated with any of $\boldsymbol{t_1^*}, \boldsymbol{t_2^*}, \ldots, \boldsymbol{t_S^*}$. We then compare the t value associated with $H_{(1)}$ to this null distribution, and reject the null hypothesis at $\alpha = 0.05$ only if this t-value exceeds 95% of the t-values in the null distribution. We then continue with the second hypothesis, however now construct our null distribution using *only* the maximum of $\boldsymbol{t_2^*}, \ldots, \boldsymbol{t_S^*}$ (ie we remove the null t-distribution associated with those variables already tested). We then follow a similar rejection procedure as above. We complete the Romano Wolf test procedure once we have tested all the hypotheses in this way, where at each stage we *only* consider the t^*-values coming from the hypotheses which have not yet been tested. Thus, at each stage the rejection criteria becomes slightly less demanding, as was the case in Holm (1979)'s procedure, but at the same time this procedure efficiently accounts for any type of correlation among the variables tested.

4.4.2 Controlling the FDR

Procedures to control for the false discovery rate came to the fore much later than those to control the family wise error rate. Nonetheless, both FDR and FWER procedures are now frequently employed. As discussed in sections above, altohugh control of the FDR allows for a small proportion of type I errors, it brings with it greater power than that available in controlling for the FWER. An extremely nice analysis of these methods in an applied context is provided by Anderson (2008) as well as a particularly elegant discussion of the types of circumstances in which we may prefer FWER or FDR corrections.[6]

The earliest suggestion of controlling for the expected proportion of falsely rejected hypotheses (the FDR) comes from Benjamini and Hochberg (1995). They propose a simple methodology, and prove that its application acts to control the FDR. They suggest the following procedure, where as above we refer to S hypothesis tests: $H_{(1)}, H_{(2)}, \ldots, H_{(S)}$, which we have ordered from most to least significant: $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(S)}$. Suppose that we define some significance level for rejection (such as 0.05) which we will denote q. Then, let k be the largest value of j for which:

$$p_{(j)} \leq \frac{j}{S}q. \tag{4.6}$$

This leads to the rejection rule to reject all $H_{(j)}$ such that $j = 1, 2, \ldots, k$, and do not reject any of the remaining hypotheses. It is important to note that this is actually a *step-up* rather than step-down procedure, as we start with the least significant hypothesis, and step up until we meet the condition in equation 4.6.

More recent methods have shown how we can improve on this first generation FDR control method (see for example the method proposed in Benjamini et al. (2006)). Nevertheless, these methods still follow the basic step-up procedure described in Benjamini and Hochberg (1995). A useful applied discussion of these various methods, as well as their implemention, can be found in Newson (2010).

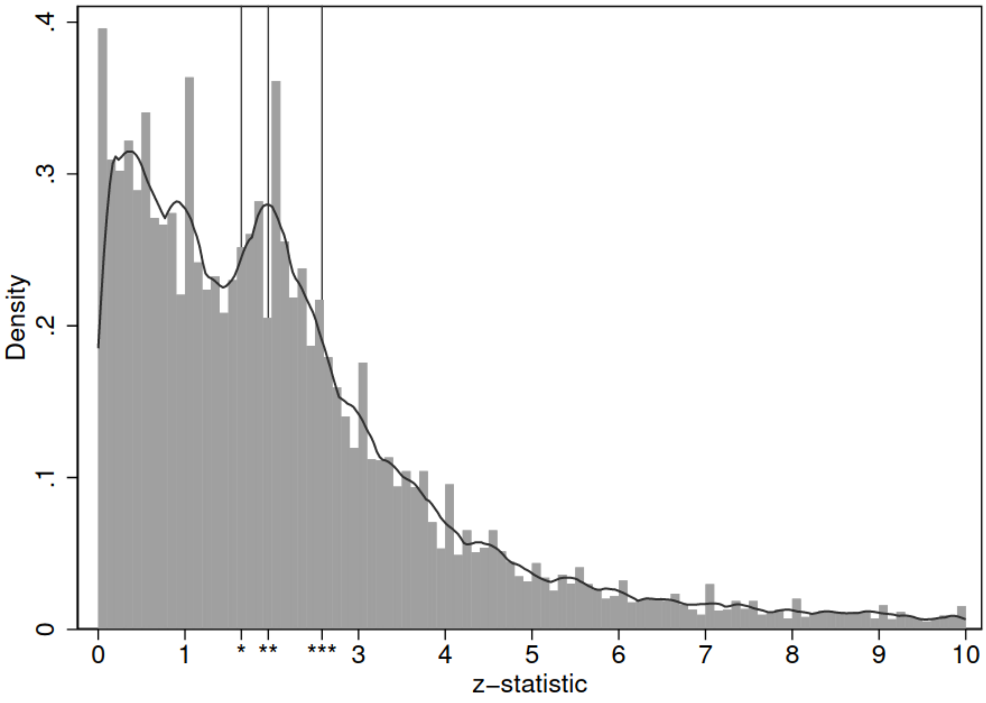
[6] This discussion, from p. 1487 of Anderson (2008) and related to assessing the impact of early childhood intervention programs is reproduced here:

> "FWER control limits the probability of making *any* type I error. It is thus well suited to cases in which the cost of a false rejection is high. In this research, for instance, incorrectly concluding that early interventions are effective could result in a large-scale misallocation of teaching resources. In exploratory analysis, we may be willing to tolerate some type I errors in exchange for greater power, however. For example, the effects of early intervention on specific outcomes may be of interest, and because overall conclusions about program efficacy will not be based on a single outcome, it seems reasonable to accept a few type I errors in exchange for greater power."

4.5 Pre-registering Trials

When working on an empirical research paper, a researcher generally faces many relatively banal choices, but which require a decision in arriving to models. To give one simple example, variables can be treated in levels or logs, and often both options may be reasonable. Provided that such decisions are justified ex-ante, it is generally perfectly fine to simply choose one or the other. However, if output to statistical models is examined in order to justify the choice made, this is problematic, as standard testing rates will break down. This is particularly concerning if there are 'rule-of-thumb' significance levels which researchers consider 'important' to advancing a particular narrative. For example, the well known significance levels of $\alpha = 0.01$, 0.05, or 0.10 may seem to be tempting targets, which of course runs entirely counter-intuitively to the nature of seeking to test a hypothesis. There is evidence from the literature that such "data snooping" procedures may occur. For example, the distribution below collated by Brodeur et al. (2020) plots Z-statistics from over 20,000 hypothesis tests in nearly 1,000 papers in empirical economics. There is evidence of important heaping of reported Z-statistics at points in the probability distribution corresponding to 'standard' significance levels, particularly $Z = 1.96$. Such a pattern would not be expected to be observed if model specifications were made without seeing statistical results. This heaping is shown to be particularly accute in certain types of empirical methods (for example IV and difference-in-difference models).

Figure 4.4: Z-statistics from empirical economic-papers



Recently, there has been growing interest in the use of pre-registered trials in the social sciences, and in experimental economics in particular (Miguel et al., 2014). The idea of pre-registering a trial is that *prior* to examining any data or running any analysis, the methodology and variables used should be entirely pre-specified, removing any concerns that specifications are chosen ex-post to fit a particular interepretation. Multiple online-registers exist including the AEA's experimental registry, where researchers can fully pre-specify their experimental hypotheses as well as their identification strategy and the precise outcome variable to be examined.

A number of suggested steps to follow when pre-registering a trial (or writing a pre-analysis plan), are laid out in Christensen and Miguel (2016). They also provide a list of noteable studies using such a plan, which are becoming much more frequent in recent literature. The use of a pre-analysis plan is particularly well-suited to an experimental study or randomised control trial in which all details can be worked out and defined *before* any data is collected. If writing a pre-analysis plan in economics, Christensen and Miguel (2016) is an excellent place to start.

Despite their growing use, a number of issues surrounding pre-analysis plans are laid out in Olken (2015). Among others, these plans may become ungainly, particularly when the design of one test is conditional on the outcome of another. Also, the extension to a non-experimental setting is not necessarily trivial. While in an experimental set-up there is a clear "before" period in which the pre-analysis plan can be written, with observational data this often is not the case. Nevertheless, and indeed as pointed out by Olken (2015), there are multiple benefits of pre-analysis plans—beyond just increased confidence in results—implying that the process of pre-specifying and registering a trial may be a valuable process to follow in many settings.

Chapter 5

Beyond Average Treatment Effects...

Required Readings
Imbens and Wooldridge (2009): Sections 3.2-3.4
Angrist and Pischke (2009): Chapter 7

Suggested Readings
Dehejia et al. (2015)
Attanasio et al. (2012)
Deaton (2010)
Heckman (2010)

5.1 The Big Picture

The methods discussed so far in this lecture series, and the literature which draws on these sorts of methods, focus very carefully on how to infer causality. Explicit questions on what drives observed outcomes—receipt of treatment, or selection into treatment—are deeply embedded in this framework. We have encountered this focus throughout all these lectures, starting from the Rubin Causal Model on the first day.

Nevertheless, it would be farfetched to suggest that our studies in economics and microeconometrics could ever be reduced simply to questions on causality, and even more farfetched to suggest that it could again be reduced only to those things that are directly manipulable by the experimenter. In the first place, the type of questions which one can ask with these methods is finite. There are many big picture questions that we would like to know about as empirical economists that can never be manipulated in an RCT, and are tricky even when thinking in terms of the wider set of methods we have discussed in other lectures.

Secondly, these methods do not lay claim to being able to respond to a question without context. When we estimate a treatment effect using these methods, it holds *only* in the context of the reform examined (that is, it is *internally valid*), and what's more, only when considering the average person subject to treatment. When we know about schooling in Mexico, or microcredit in Pakistan, or worms in Kenya, do we know anything about these problems in other places in the world? In a strict econometric sense, no. The treatment effects literature (and particularly RCTs), receive critique for having a lack of *external validity*, meaning that what we learn in one context will not necessarily hold in another.[1]

There are many papers which debate the merits of reduced form work like that described here, and more extensive econometric methods, including structural modelling. The suggested reading by Deaton is a very good place to start. While we only provide a brief introduction in these lectures, it is worth noting a couple of things in closing: firstly about methods which increase the scope of these results, and secondly about if (and if so, how) these results tie into the wider world of structural econometrics.

5.2 Heterogeneity and Quantile Treatment Effects

Firstly, we will consider how the *heterogeneity* of individuals can tie in with these methods. In general, in our econometrics up to this point, we have been content to estimate an average parameter—for example $\hat{\beta}$ estimated by OLS, or some other type of *average* treatment effect—however we have not thought too far beyond these average responses. At times, the average effect of a reform may be truly what we would like to know. However other times, it certainly won't be. For example, consider a program targeted to schooling outcomes. If a large average treatment effect is driven only by those with very high test scores, we may consider that the program is actually not doing a good job in addresing problems in the true population of interest, such as children at risk of not progressing, or learning key skills. Indeed, we may be particularly interested with just a certain group of the population, such as those in the bottom half or bottom quartile of schooling outcomes, if we are aiming to avoid particularly poor results. In any case, and in general, there are certainly considerations of equality which remain hidden when an average treatment effect is reported.

[1] It is interesting to think of the parallel between economics and the natural sciences in this case. When something is demonstrated in a laboratory in the natural sciences, typically it is done so under standardised conditions (such as "Standard Temperature and Conditions" (STC). In this sense, external validity is not important, as these conditions can be replicated anywhere, and the validity of the result can then be proved. In development economics, there is no such thing as STC! The local conditions and institutions in one country, region, village etc., do not exist in other places. As such, a positive result in one circumstance, while at least providing proof of concept, tells us very little about what would actually happen were the same policy to be implmented in a different context.

5.2.1 An Introduction to Quantile Regressions

A simple way to unpack heterogeneity in a regression framework is by estimating a quantile regression. The quantile regression allows for the calculation of the impact of some variable (or variables) x on some dependent variable of interest y at different quantiles of the dependent variable y. Thus, rather than estimate a single β capturing the impact of x on y, we can estimate a series of β_q, capturing the impact of x on y at percentile q of the variable y. These percentiles may be the median, or quartile 1 (the lowest 25% of the population), or any percentile of interest. One important thing to note is that we are *not* referring to different percentiles of the independent variables x, but rather examining how outcomes vary across the distribution of the dependent variable. Below we provide an example where the dependent variable of interest is birth weight. In this case, for example, a quantile regression of a baby's birthweight (y) on the number of cigarettes that a mother smokes during pregnancy (x) would provide a different estimate for each quantile of birthweight, *not* for different quantiles of cigarette consumption.

In practice, quantile regression parameters are estimated for a particular percentile, which we call q, and, as long as heterogeneity is present, will vary for each q. It is common for the parameters to be displayed at a range of quantiles, for example as documented in the graphs in Figure 5.1. The estimation procedure in quantile regression is via an absolute error loss function (rather than a squared error loss function, as in OLS). As a result, quantile regression is less sensitive to outliers than OLS. In particular, the quantile regression estimator $\hat{\beta}_q$ for percentile q minimizes the following loss function:

$$Q_N(\beta_q) = \sum_{i:y_i \geq x_i'\beta} q|y_i - x_i'\beta_q| + \sum_{i:y_i < x_i'\beta} (1-q)|y_i - x_i'\beta_q|.$$

Note that here we are ordering the dependent variable y_i, and consider the observations above the median in the left-hand term, and below the median in the right hand term. Also note that we are "tilting" the optimization in favour of the lower or upper percentiles depending on the value of q. If q is a high percentile, we will give more weight to the left-hand term, and so $\hat{\beta}_q$ will take more into account information on observations above the median, and vice versa. In the special case of the median ($q = 0.5$), this regression simply collapses to the least absolute deviation estimator. Additional details of this estimator can be found in Cameron and Trivedi (2005, section 4.6) and Angrist and Pischke (2009, chapter 7).

Additionally, a brief introduction to the quantile regression can be found in Koenker and Hallock (2001) who document an example using US data on birth weights. Below we replicate their empirical example, however using more recent data on all birth weights in the USA in 2015. This is a quantile regression where the dependent variable is each baby's birthweight, and with 11 independent variables (plus a constant). The quantile regression estimates $\hat{\beta}_q$ are documented for each of the independent variables, and at each percentile of the distribution of birth weight. These are presented along with their standard errors (in grey), and the parameters from a linear OLS regression (dashed lines). Note for example the interpretation of the variable "smoker" (bottom left panel). While the mean impact of smoking on birthweight is around a 150 gram reduction, this impact is *largest* in the lowest birth weight quantiles, suggesting that smoking is particularly damaging for babies which are already at a very low birthweight. While we will not discuss the full range of coefficients, notice that there is no impediment to using both continuous and discrete variables in these models.

5.2.2 Quantile Treatment Effects

We will begin by thinking about generalizing average treatment effects to treatment effects at different points of the outcome distribution. This is very closely related to the quantile regressions discussed above, but in particular, cast in the treatment effect framework. As discussed above, the idea of a quantile regression is to look at the estimated effect of some variable x at different points of the y distribution. The Quantile Treatment effect (QTE) for quantile q is the effect of the treatment evaluated at the quantile q (e.g. the 10th percentile, the median, the 90th percentile,...) of the distribution of the outcome variable. Let us denote by τ_q the QTE for quantile q, so that we have:

$$y_i = \tau_q T_i + u_i$$

If the treatment is binary, just as the ATE was the difference in mean outcome between treatment and control, the $QTE(q)$ is the difference between the quantiles q of the distribution of outcome in treatment and control. Let us denote by F_{Y_1} and F_{Y_0} the distribution function of outcomes in treatment and control respectively:

$$\tau_q = F_{Y_1}^{-1}(q) - F_{Y_0}^{-1}(q).$$

It is worth noting briefly that an important difference between this and the ATE framework is that $QTE(q)$ will, in general, be different from the quantile of the differences in outcomes between treatment and control. If we denote by $F_{Y_1-Y_0}$ the distribution function of the difference in outcomes:

$$\overline{\tau_q} = F_{Y_1-Y_0}^{-1}(q)$$

In other terms "the quantile of differences is not the difference of the quantiles". If you cast your mind back to non-linear regression models, this may remind you somewhat of the difference between average marginal effects of a probit or logit, versus the marginal effect at the mean. Graphically, what we are trying to capture with a quantile treatment effect at different parts of the distribution is displayed in figure 5.2.

Let's consider the case of an RCT with perfect compliance. In that case the treatment is

Figure 5.1: Quantile Regression and Birth Weight

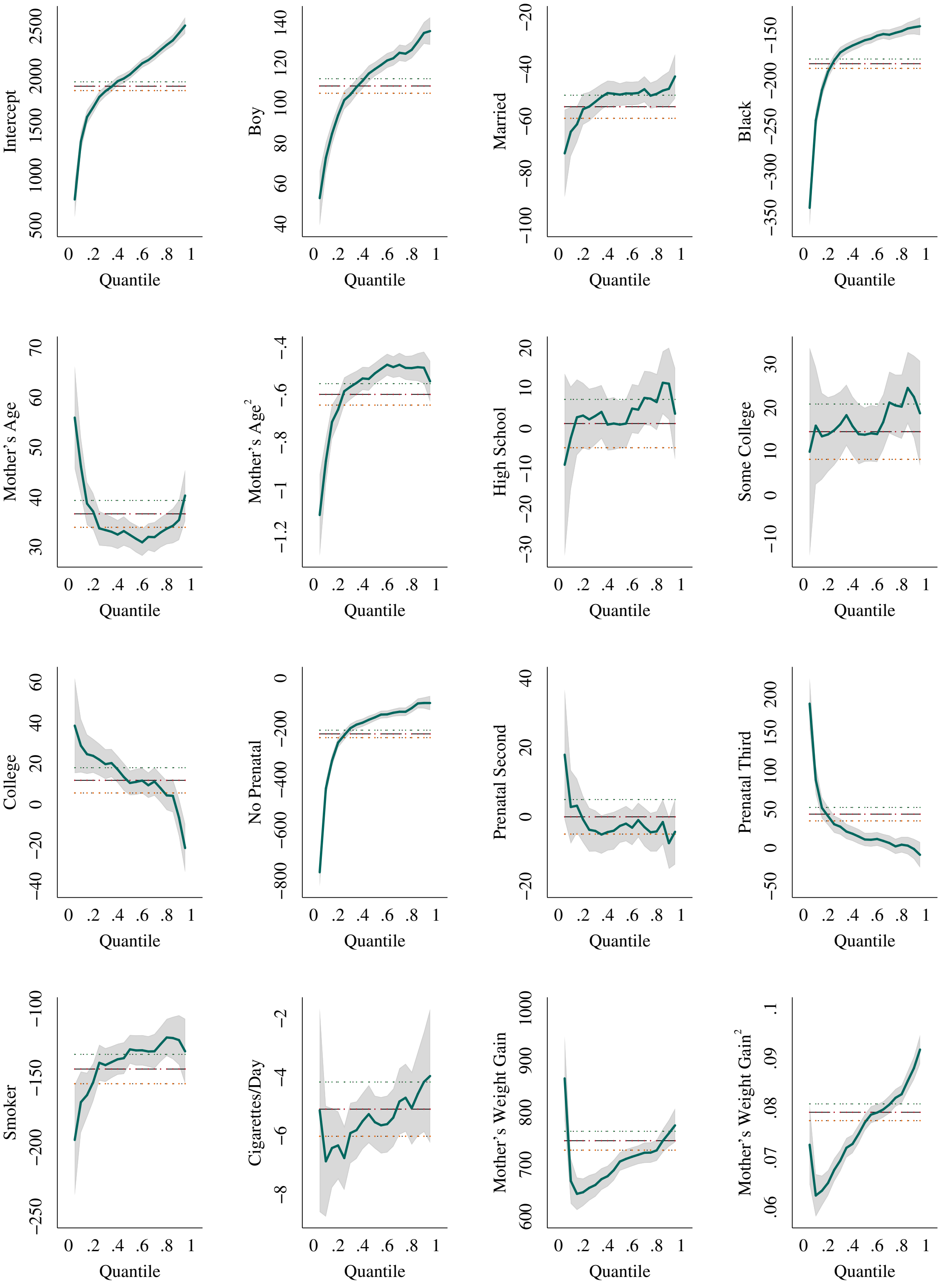
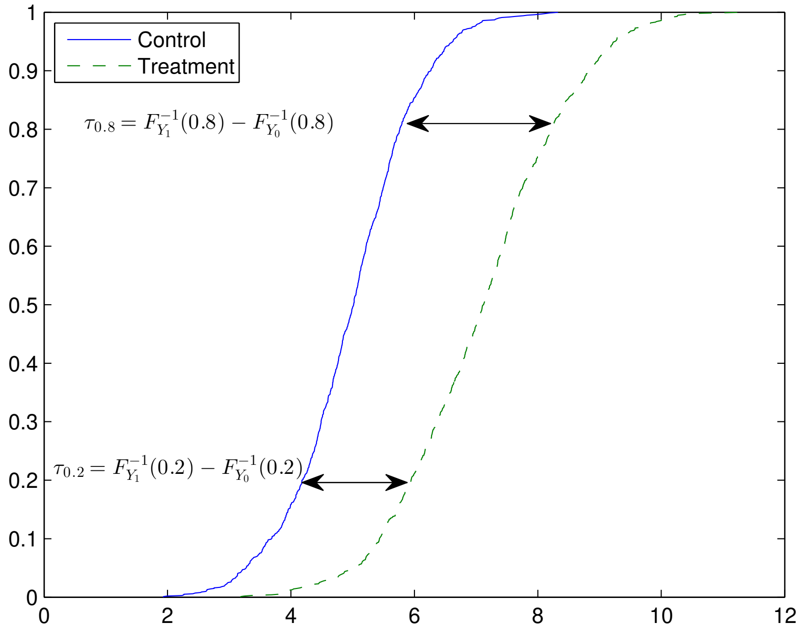


Figure 5.2: Quantile treatment effects



exogenous with respect to potential outcomes, we can consistently estimate the $QTE(q)$ by estimating τ_q through quantile regressions of the form:

$$y_i = \tau_q T_i + \varepsilon_i$$

Quantile regressions use information which we previously threw out of the estimation when estimating the ATE with OLS at the start of this lecture series. The heterogeneity in treatment responses was previously treated only as an issue of heteroscedasticity. For example, if we assume that the treatment effect increases linearly with the outcome then:

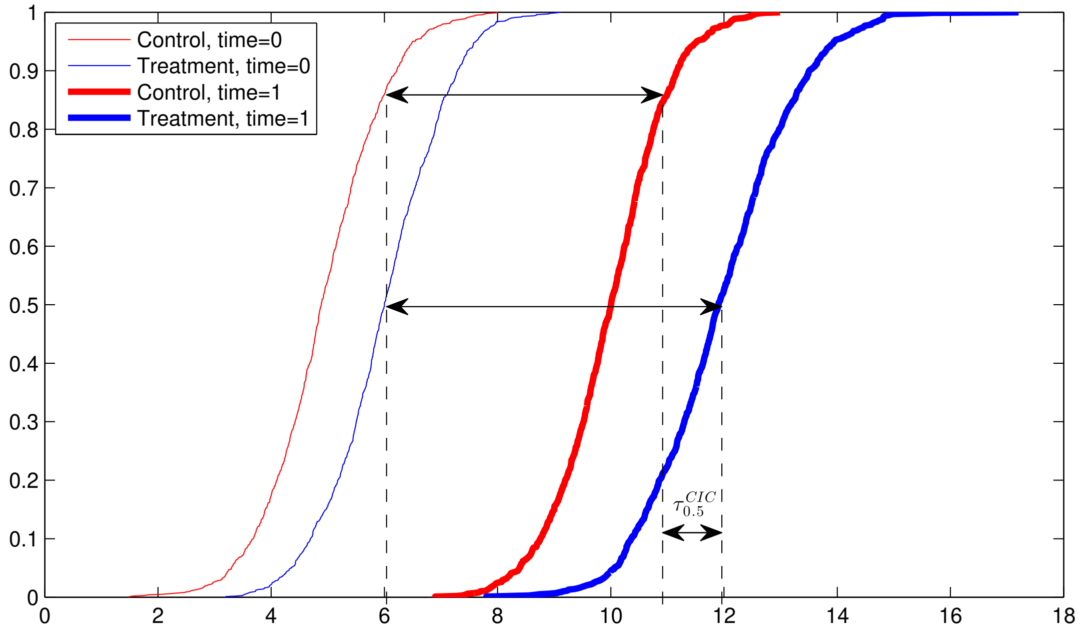
$$y_i = \mu_0 + (\mu_1 - \mu_0)T_i + (\gamma T_i e_i + e_i)$$

So that OLS estimates were consistent but the variance of the error term $\varepsilon_i = (\gamma T_i e_i + e_i)$ increased with treatment. The distributions displayed in figure 5.2 display behaviour of this type. In this situation, our average tretment effect is hiding some important, and potentially very policy relevant, heterogeneity!

5.2.3 Changes-in-Changes

Athey and Imbens (2006) have proposed a way to use quintiles to implement a strategy close to DiD. Let us assume that we observe two cross-sections of the treatment group and control group at two periods, one before (t) and one after the treatment ($t + 1$).

Figure 5.3: Change in Change



They idea is to match at time t each quintile q in the treatment group with quintile q' in the control group with the same value of outcome, and then compare the change in outcomes for quintile q to changes in outcomes for quintile q' (see Figure 5.3).

Let us denote F_{gt} by the distribution functions of the outcome for group g (0 for control and 1 for treatment) at time t. The change in change estimator is equal to:

$$ATT^{CIC} = \frac{1}{N_1} \sum_{i \in G_1} y_{it+1} - F_{0t+1}^{-1}(F_{0t}(F_{1t}^{-1}(F_{1t+1}(y_{it+1}))))$$

The change in change is the ATT under two conditions: monotonicity, i.e. treatment does not change the rank and conditional independence (as for DiD methods).

Although the focus of Athey and Imbens (2006) is on developing a new method to estimate average treatment effect, their method also allows to consider the effect of treatment at different quintiles of the distribution of earnings:

$$\tau_q^{CIC} = F_{1t+1}^{-1}(q) - F_{0t+1}^{-1}(F_{0t}(F_{1t}^{-1}(q)))$$

5.3 Treatment Effects and External Validity

These methods, while allowing us to examine heterogeneity *within* the reduced scope of one particular program, still provide no guidance on whether a result in a particular *context* is applicable in another location. Recent work is starting to think about these questions in a more formal way.

One particular approach is suggested by Dehejia et al. (2015) who suggest an "external validity function", which asks how far an experiment run in a particular context may be from the mean effect in all locations. They examine the estimated effect of an additional child on his or her mother's labour supply in many contexts. While we will not go into the technical details here, if this is relevant for your work, I encourage you to consult the suggested reading.

5.4 Sorting

When we spoke about the LATE we first began to think about sorting. In the case of assignment to treatment and the LATE we allowed a certain type of sortig whereby individuals could choose whether to comply with a randomly assigned treatment status. However, more generally, sorting is a pervasive issue which we must deal with when considering economic and econometric models. Logically, in most cases, individuals make decisions based on what they perceive is best for them. Or in other words, when possible, individuals will "sort" themselves based on their potential outcomes.

Perhaps the most well known sorting model in labour economics is the Roy Model. This is a simple model where individuals choose their 'treatment status' based on their outcomes under two scenarios. In an econometric sense, we can think of this as a decision made completely endogenously to the system under study. We will begin this section by laying out the Roy Model as a general framework to think about a broader class of sorting decisions, before turning to a particular model of thinking about treatment effects under endogenous decisions and sorting: the Marginal Treatment Effects framework.

5.4.1 The Roy Model

The classic Roy Model is laid out in an expositional paper (Roy, 1951) entitled "Some Thoughts on the Distribution of Earnings". The paper itself lays out all the details of the model without formally writing it down, sketching a clear picture of selection into (two) occupations based on the potential rewards which each person faces in each occupation. Specifically, the paper speaks of a village where individuals decide whether to fish or hunt rabbits, though abstractly the model applies to any cases where individuals seek to consider their wellbeing in both states of a decision when deciding between two options. In this sense, it is clear how it ties in with the "potential outcomes" framework we are using in this course: here individuals will select into the "treatment status" which is most beneficial to them given their particular payoff. To the degree that an econometrician does not observe the payoffs an individual perceives to both states of the world, their will be challenges in identifying casual effects. Here I briefly layout out the Roy Model as a way to explicitly think about selection, before turning to a more general set-up which considers estimation and identification where individuals decide upon their treatment status.

The Roy model lays out a simplified framework where each individual freely chooses to fish or hunt rabbits (exclusively), and the market pays a price for fish denoted π_F and a price for rabbits denoted π_R. Thus, an individual who catches F_i fish if they choose to fish or R_i rabbits if they choose to hunt would receive a wage of:

$$W_{Fi} = \pi_F F_i \tag{5.1}$$

$$W_{Ri} = \pi_R R_i. \tag{5.2}$$

Assuming for simplicity there is no uncertainty in these quantities, an individual would choose to fish if $W_{Fi} > W_{Ri}$. Roy (1951) states that hunting is easier, whereas fishing requires more skill. This is quite a simple model, but can be extended in many ways (see for example a summary in Heckman and Taber (2010))

Individuals have different skill levels, which is to say that there is heterogeneity in F_i and R_i. Specifically, Roy (1951) assumes that the log of "skills" (the level of production of each good) are jointly normally distributed:

$$\begin{bmatrix}\log(F_i)\\ \log(R_i)\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\mu_F\\ \mu_R\end{bmatrix}, \begin{bmatrix}\sigma_F^2 & \rho_{FR}\\ \rho_{FR} & \sigma_R^2\end{bmatrix}\right). \tag{5.3}$$

A key goal of the Roy model is understanding self-selection. Into which tasks will the most efficient workers select based on this structure? Note that so far, we have not assumed any particular type of correlation between skills in hunting and fishing (that is to say, we have not assumed a sign for ρ_{FR}). To determine whether more efficient workers are selected into particular tasks, say fishing, we need to determine whether $E[\log(F_i)|\pi_F F_i > \pi_R R_i]$ is greater than $E[\log(F_i)]$, that is to say, whether the average salary of individuals who fish is above the average salary of all potential individuals if they would fish.

It can be shown (see Heckman and Taber (2010, p. 222))[2] that the value of this conditional

[2] We won't go through the notation here, though if you would like to see how this is resolved, Chris Taber has some slides laying this out quite extensively. These are available at: `https://www.ssc.wisc.edu/~ctaber/751/roy.pdf`, slides 22-29 are the most relevant for this calculation.

expectation can be written as:

$$E[\log(F_i)|\pi_F F_i > \pi_R R_i] = \mu_F + \frac{(\sigma_F^2 - \rho_{FR})}{\sigma}\lambda\left(\frac{\log(\pi_F) - \log(\pi_R) + \mu_F - \mu_R}{\sigma}\right), \quad (5.4)$$

where σ^2 is the variance of log(F_i/R_i), and $\lambda(\cdot)$ is the inverse Mills ratio, which is found in cases where a truncated normal distribution is considered. The important thing to note here is that this expectation is equal to the mean of $log(F_i)$ (the average skill in the population), plus a second term, which describes the nature of selection. Thus, if this second term is positive, this implies more skilled individuals select into fishing, while if it is negative, less-skilled (in fishing) individuals select into fishing. Note that the inverse Mills ratio is always positive, and standard deviation σ must be positive. So the nature of selection depends entirely on the sign of $(\sigma_F^2 - \rho_{FR})$. Note also that $\sigma^2 = (\sigma_F^2 - \rho_{FR}) + (\sigma_R^2 - \rho_{FR}) > 0$, so one of the two terms must be positive (and both can be positive), implying positive selection into at least one of fishing or hunting. Based on all this, a number of general results can be summarised:

- If fishing is harder and there is a larger variance in fishing ability in the population $\sigma_F^2 > \sigma_R^2$ which must imply positive selection into fishing.

- In the case of hunting (the lower variance occupation), the nature of selection depends on the value of ρ_{FR} relative to σ_R^2.

  - If ρ_{FR} (hunting skill and fishing skill are negatively correlated), there will be positive selection into hunting too
  - If hunting and fishing skill are perfectly correlated, given that $\sigma_F^2 > \sigma_R^2$, then ρ_{FR} must be larger than σ_R^2, and there will be negative selection into hunting
  - For cases in between negative correlations and perfect positive correlations, either case can arise.

This simplified model thus already gives some framework to think about selection and how a population of individuals will behave if they are seeking to maximise payoffs among choices. There are many extensions to these models, and applications where it is used as a basis for estimation with real data (for example Taber and Vejlin (2020)). However, here we lay out the Roy model as a precursor for thinking about heterogeneity, given its importance in thinking about marginal treatment effects, and the value for particular individuals of selecting into one or other case. In the case laid out here, certain individuals may have a much higher wage in one of two occupations and hence have much to gain from choosing this, while others may have reasonably similar wages in both occupations, and thus less to gain from their occupational choice. This is something we turn to examine now.

5.4.2 "Marginal Treatment Effects" and Other Relevant Quantities

The Marginal Treatment Effect (MTE) framework starts with a more flexible version of a Roy-style model. Heckman and Vytlacil (2005), who formalize the marginal treatment effects framework, refer to the generalized Roy model, which augments the above standard Roy model with a component capturing the cost of receiving treatment. We now define Z as observables cost of receiving treatment, such that an individual would select into treatment if their benefits of treatment exceed the benefits of not electing to receive treatment, net of any costs. The interest of this model is in allowing very flexibly for selection into treatment, and considering which treatment effects can, and ideally should, be estimated. At its heart, the MTE framework is about selection into treatment, and so begins with a consideration of treatment itself:

$$D^* = \mu_D(Z,X) - U_D, \qquad D = 1 \text{ if } D^* \geq 0, \;\text{ else }\; D = 0 \tag{5.5}$$

Note that this views selection into treatment as a latent variable model, similar to latent variables underlying standard binary choice models such as the probit or logit. Underlying these terms, are potential outcomes:

$$Y_1 = \mu_1(X, U_1) \qquad Y_0 = \mu_0(X, U_0)$$

and costs of receiving treatment

$$C = \mu_c(Z) + U_C.$$

Based on this, we can understand how an individual will make their choices – their D^* will be greater than or equal to 0, implying selection into treatment, if $Y_1 - Y_0 - C \geq 0$. Thus, note above in equation 5.5 that in the generalized Roy model setting, $U_D = U_1 - U_0 - U_C$, which all refer to unobservable terms.

The use of Z in relying to costs is not casual. In Heckman and Vytlacil (2005), Z is assumed to affect the likelihood of opting into treatment, while also being independent of U_1, U_0 and U_D. Thus, these "cost of treatment" components are actually viewed as an instrumental variable. This would be particularly clear if Z was a random assignment to treatment, and brings us back to the setting described in section 3.1 when discussing IV and the LATE. Where Heckman and Vytlacil (2005) seek to go considerably beyond LATE, however, is in formally modelling selection into treatment, and thinking about the resulting estimands. They define $P(Z)$ as the probability of receiving treatment given any particular vaue of Z, or $P(Z) \equiv Pr(D = 1|Z)$. This is, of course, just a propensity score, given that it relates the likelihood of receiving treatment with some observable characteristic(s).

Based on the above definitions,

UNDER CONSTRUCTION. THIS SECTION WILL BE UPDATED DURING SEMESTER.

Bibliography

A. Abadie. Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics*, 113(2):231–263, 2003.

A. Abadie and J. Gardeazabal. The Economic Costs of Conflict: A Case Study of the Basque Country. *American Economic Review*, 93(1):113–132, 2003.

A. Abadie, A. Diamond, and J. Hainmueller. Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program. *Journal of the American Statistical Association*, 105(490):493–505, 2010.

D. Almond. Is the 1918 Influenza Pandemic Over? Long-Term Effects of In Utero Exposure in the Post-1940 U.S. Population. *Journal of Political Economy*, 114(4):672–712, 2006.

M. L. Anderson. Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects. *Journal of the American Statistical Association*, 103(484):1481–1495, 2008.

J. Angrist, V. Lavy, and A. Schlosser. Multiple Experiments for the Causal Link between the Quantity and Quality of Children. *Journal of Labor Economics*, 28(4):773–824, October 2010.

J. D. Angrist and G. W. Imbens. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American Statistical Association*, 90(430):431–442, 1995. doi: 10.1080/01621459.1995.10476535.

J. D. Angrist and V. Lavy. Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement. *The Quarterly Journal of Economics*, 114(2):533–575, 1999.

J. D. Angrist and J.-S. Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, 2009.

O. Ashenfelter. Estimating the Effects of Training Programs on Earnings. *Review of Economics and Statistics*, 60(1):47–57, 1978.

S. Athey and G. Imbens. Chapter 3 - the econometrics of randomized experimentsa. In A. V. Banerjee and E. Duflo, editors, *Handbook of Field Experiments*, volume 1 of *Handbook of Economic Field Experiments*, pages 73 – 140. North-Holland, 2017. doi: https://doi.org/10.1016/bs.hefe.2016.10.003. URL http://www.sciencedirect.com/science/article/pii/S2214658X16300174.

S. Athey and G. W. Imbens. Identification and Inference in Nonlinear Difference-in-Differences Models. *Econometrica*, 74(2):431–497, 2006.

S. Athey and G. W. Imbens. Design-based Analysis in Difference-In-Differences Settings with Staggered Adoption. NBER Working Papers 24963, National Bureau of Economic Research, Inc, Aug. 2018. URL https://ideas.repec.org/p/nbr/nberwo/19305.html.

O. P. Attanasio, C. Meghir, and A. Santiago. Education choices in mexico: Using a structural model and a randomized experiment to evaluate progresa. *The Review of Economic Studies*, 79(1):37–66, 2012.

S. Baird, J. H. Hicks, E. Miguel, and M. Kremer. Worms at Work: Long-run Impacts of a Child Health Investment. *Quarterly Journal of Economics*, 131(4):1637–1680, Jul 2016.

A. V. Banerjee and E. Duflo. The Experimental Approach to Development Economics. *Annual Review of Economics*, 1(1):151–178, 05 2009. URL https://ideas.repec.org/a/anr/reveco/v1y2009p151-178.html.

V. Baranov, S. Bhalotra, P. Biroli, and J. Maselko. Maternal depression, women's empowerment, and parental investment: Evidence from a randomized controlled trial. *American Economic Review*, 110(3):824–59, March 2020. doi: 10.1257/aer.20180511. URL https://www.aeaweb.org/articles?id=10.1257/aer.20180511.

L. Beaman, E. Duflo, R. Pande, and P. Topalova. Female Leadership Raises Aspirations and Educational Attainment for Girls: A Policy Experiment in India. *Science*, 335(6068):582–586, 2012.

Y. Benjamini and Y. Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

Y. Benjamini, A. M. Krieger, and D. Yekutieli. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, 93(3):491–507, 2006.

M. Bertrand, E. Duflo, and S. Mullainathan. How Much Should We Trust Differences-In-Differences Estimates? *The Quarterly Journal of Economics*, 119(1):249–275, 2004.

S. Bhalotra and D. Clarke. The Twin Instrument: Fertility and Human Capital Investment. *Journal of the European Economic Association*, 12 2019a. ISSN 1542-4766. doi: 10.1093/jeea/jvz058. URL https://doi.org/10.1093/jeea/jvz058. jvz058.

S. Bhalotra and D. Clarke. Twin birth and maternal condition. *The Review of Economics and Statistics*, 101(5):853–864, 2019b.

S. Bhalotra, D. Clarke, J. F. Gomes, and A. Venkataramani. Maternal Mortality and Women's Political Participation. CEPR Discussion Papers 14339, C.E.P.R. Discussion Papers, Jan. 2020. URL https://ideas.repec.org/p/cpr/ceprdp/14339.html.

P. Bharadwaj, K. V. Løken, and C. Neilson. Early Life Health Interventions and Academic Achievement. *American Economic Review*, 103(5):1862–1891, 2013.

H. S. Bloom. Accounting for No-Shows in Experimental Evaluation Designs. *Evaluation Review*, 8(2):225–246, 1984.

C. E. Bonferroni. Il calcolo delle assicurazioni su gruppi di teste. In *Studi in Onore del Professore Salvatore Ortu Carboni*, pages 13–60. Rome, 1935.

K. Borusyak and X. Jaravel. Revisiting Event Study Designs, with an Application to the Estimation of the Marginal Propensity to Consume. mimeo, 2018.

A. Brodeur, N. Cook, and A. Heyes. Methods matter: p-hacking and publication bias in causal analysis in economics. *American Economic Review*, 110(11):3634–60, November 2020. doi: 10.1257/aer.20190687. URL https://www.aeaweb.org/articles?id=10.1257/aer.20190687.

F. Brollo and U. Troiano. What happens when a woman wins an election? Evidence from close races in Brazil. *Journal of Development Economics*, 122(C):28–45, 2016.

B. Callaway and P. H. Sant'Anna. did: Treatment Effects with Multiple Periods and Groups. Comprehensive R Archive Network, Feb. 2020. URL https://cran.r-project.org/web/packages/did/index.html.

B. Callaway and P. H. Sant'Anna. Difference-in-differences with multiple time periods. *Journal of Econometrics*, 2021. ISSN 0304-4076. doi: https://doi.org/10.1016/j.jeconom.2020.12.001. URL https://www.sciencedirect.com/science/article/pii/S0304407620303948.

S. Calonico, M. D. Cattaneo, and R. Titiunik. Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs. *Econometrica*, 82(6):2295–2326, 2014a.

S. Calonico, M. D. Cattaneo, and R. Titiunik. Robust data-driven inference in the regression-discontinuity design. *The Stata Journal*, 14(4):909–946, 2014b.

S. Calonico, M. D. Cattaneo, and R. Titiunik. Optimal Data-Driven Regression Discontinuity Plots. *Journal of the American Statistical Association*, 110(512):1753–1769, 2015. doi: 10.1080/01621459.2015.1017578.

A. C. Cameron and D. L. Miller. A practitioner's guide to cluster-robust inference. *The Journal of Human Resources*, 50(2):317–72, 2015.

A. C. Cameron and P. K. Trivedi. *Microeconometrics: Methods and Applications*. Cambridge University Press, 2005.

A. C. Cameron, J. B. Gelbach, and D. L. Miller. Bootstrap-Based Improvements for Inference with Clustered Errors. *Review of Economics and Statistics*, 90(3):414–427, 2008.

D. Card, D. S. Lee, Z. Pei, and A. Weber. Inference on Causal Effects in a Generalized Regression Kink Design. *Econometrica*, 83(6):2453–2483, 2015.

P. Carneiro, J. J. Heckman, and E. Vytlacil. Evaluating marginal policy changes and the average effect of treatment for individuals at the margin. *Econometrica*, 78(1):377–394, 2010.

G. Casella and R. L. Berger. *Statistical Inference*. Duxberry Thomson, 2 edition, 2002.

M. D. Cattaneo and R. Titiunik. Regression Discontinuity Designs. *Annual Review of Economics*, 14:1–48, 2022.

G. S. Christensen and E. Miguel. Transparency, reproducibility, and the credibility of economics research. Working Paper 22989, National Bureau of Economic Research, December 2016.

D. Clarke and K. Tapia Schythe. Implementing the panel event study. IZA Discussion Papers 13524, Institute of Labor Economics (IZA), 2020.

D. Clarke, S. Oreffice, and C. Quintana-Domeque. The Demand for Season of Birth. Working Papers 2016-032, Human Capital and Economic Opportunity Working Group, Dec. 2016.

I. Clots-Figueras. Are female leaders good for education? evidence from india. *American Economic Journal: Applied Economics*, 4(1):212–44, 2012.

J. Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Academic Press, 1988.

T. G. Conley and C. R. Taber. Inference with "difference in differences" with a small number of policy changes. *The Review of Economics and Statistics*, 93(1):113–125, 2011.

C. Davey, A. M. Aiken, R. J. Hayes, and J. R. Hargreaves. Re-analysis of health and educational impacts of a school-based deworming programme in western Kenya: a statistical replication of a cluster quasi-randomized stepped-wedge trial. *International Journal of Epidemiology*, 2015. doi: 10.1093/ije/dyv128.

C. de Chaisemartin and X. D'Haultfoeuille. Fuzzy Differences-in-Differences. *The Review of Economic Studies*, 85(2):999–1028, 08 2017. ISSN 0034-6527. doi: 10.1093/restud/rdx049.

C. de Chaisemartin and X. D'Haultfoeuille. Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, 110(9):2964–96, September 2020. doi: 10.1257/aer.20181169. URL https://www.aeaweb.org/articles?id=10.1257/aer.20181169.

C. de Chaisemartin, X. D'Haultfoeuille, and A. Deeb. TWOWAYFEWEIGHTS: Stata module to estimate the weights and measure of robustness to treatment effect heterogeneity attached to two-way fixed effects regressions. Statistical Software Components, Boston College Department of Economics, Feb. 2019a.

C. de Chaisemartin, X. D'Haultfoeuille, and Y. Guyonvarch. DID_MULTIPLEGT: Stata module to estimate sharp Difference-in-Difference designs with multiple groups and periods. Statistical Software Components, Boston College Department of Economics, May 2019b.

A. Deaton. *The Analysis of Household Surveys – A Microeconometric Approach to Development Policy*. The Johns Hopkins University Press, 1997.

A. Deaton. Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic development. Working Paper 14690, National Bureau of Economic Research, January 2009.

A. Deaton. Instruments, randomization, and learning about development. *Journal of Economic Literature*, 48(2):424–55, 2010.

A. Deaton. Randomization in the tropics revisited: a theme and eleven variations. Working Paper 27600, National Bureau of Economic Research, July 2020.

R. Dehejia, C. Pop-Eleches, and C. Samii. From local to global: External validity in a fertility natural experiment. NBER Working Papers 21459, National Bureau of Economic Research, Inc, 2015.

R. H. Dehejia and S. Wahba. Propensity Score-Matching Methods For Nonexperimental Causal Studies. *The Review of Economics and Statistics*, 84(1):151–161, February 2002.

M. Dell. Trafficking Networks and the Mexican Drug War. *American Economic Review*, 105 (6):1738–17792, 2015.

J. J. Diaz and S. Handa. An Assessment of Propensity Score Matching as a Nonexperimental Impact Estimator: Evidence from Mexico's PROGRESA Program. *Journal of Human Resources*, XLI(2):319–345, 2006.

W. S. Dobbie and R. G. Fryer. The medium-term impacts of high-achieving charter schools. *Journal of Political Economy*, 123(5):985–1037, 2015.

N. Doudchenko and G. W. Imbens. Balancing, Regression, Difference-In-Differences and Synthetic Control Methods: A Synthesis. Working Paper 22791, National Bureau of Economic Research, October 2016.

E. Duflo. Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment. *American Economic Review*, 91(4):795–813, September 2001.

E. Duflo, R. Glennerster, and M. Kremer. Chapter 61 using randomization in development economics research: A toolkit. In T. P. Schultz and J. A. Strauss, editors, *Handbook of Development Economics*, volume 4 of *Handbook of Development Economics*, pages 3895 – 3962. Elsevier, 2007. doi: https://doi.org/10.1016/S1573-4471(07)04061-2. URL http://www.sciencedirect.com/science/article/pii/S1573447107040612.

B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1): 1–26, 1979.

R. A. Fisher. *Statistical Methods for Research Workers*. Oliver & Boyd, 1925.

R. A. Fisher. *The Design of Experiments*. Oliver & Boyd, 1935.

S. Freyaldenhoven, C. Hansen, and J. M. Shapiro. Pre-event trends in the panel event-study design. *American Economic Review*, 109(9):3307–38, September 2019. doi: 10.1257/aer.20180609. URL http://www.aeaweb.org/articles?id=10.1257/aer.20180609.

T. Fujiwara and L. Wantchekon. Can informed public deliberation overcome clientelism? Experimental evidence from Benin. *American Economic Journal: Applied Economics*, 5(1): 241–255, 2013.

P. Ganong and S. Jäger. A Permutation Test and Estimation Alternatives for the Regression Kink Design. IZA Discussion Papers 8282, Institute for the Study of Labor (IZA), June 2014.

A. Gelman and G. Imbens. Why High-Order Polynomials Should Not Be Used in Regression Discontinuity Designs. *Journal of Business & Economic Statistics*, 37(3):447–456, 2019. doi: 10.1080/07350015.2017.1366909.

A. Gelman and E. Loken. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. Nov. 2013.

P. Gertler, J. Heckman, R. Pinto, A. Zanolini, C. Vermeersch, S. Walker, S. Chang, and S. Grantham-McGregor. Labor market returns to an early childhood stimulation intervention in Jamaica. *Science*, 344(xxxx):998–1001, 2014.

D. O. Gilligan and J. Hoddinot. Is There Persistence in the Impact of Emergency Food Aid? Evidence on Consumption, Food Security, and Assets in Ethiopia. *American Journal of Agricultural Economics*, 89(2):225–242, 2007.

R. Glennerster and K. Takavarasha. *Running Randomized Evaluations: A Practical Guide*. Princeton University Press, 2013.

A. Goodman-Bacon. Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, 2021. ISSN 0304-4076. doi: https://doi.org/10.1016/j.jeconom.2021.03.014. URL https://www.sciencedirect.com/science/article/pii/S0304407621001445.

A. Goodman-Bacon, T. Goldring, and A. Nichols. BACONDECOMP: Stata module to perform a Bacon decomposition of difference-in-differences estimation. Statistical Software Components, Boston College Department of Economics, July 2019. URL https://ideas.repec.org/c/boc/bocode/s458676.html.

C. W. J. Granger. Investigating Causal Relations by Econometric Models and Cross-Spectral Methods. *Econometrica*, 37(3):424–38, July 1969.

J. Heckman, S. Urzua, and E. Vytlacil. Understanding instrumental variables in models with essential heterogeneity. *Review of Economics and Statistics*, 88(3):389–432, 2006.

J. J. Heckman. Building bridges between structural and program evaluation approaches to evaluating policy. *Journal of Economic Literature*, 48(2):356–98, 2010.

J. J. Heckman and J. A. Smith. The Pre-programme Earnings Dip and the Determinants of Participation in a Social Programme. Implications for Simple Programme Evaluation Strategies. *The Economic Journal*, 109(457):313–348, 1999.

J. J. Heckman and C. Taber. *Roy model*, pages 221–228. Palgrave Macmillan UK, London, 2010. ISBN 978-0-230-28081-6. doi: 10.1057/9780230280816_27. URL https://doi.org/10.1057/9780230280816_27.

J. J. Heckman and E. Vytlacil. Structural equations, treatment effects, and econometric policy evaluation. *Econometrica*, 73(3):669–738, 2005. doi: 10.1111/j.1468-0262.2005.00594.x.

S. Heß. Randomization inference with stata: A guide and software. *The Stata Journal*, 17(3):630–651, 2017.

J. H. Hicks, M. Kremer, and E. Miguel. Commentary: Deworming externalities and schooling impacts in Kenya: a comment on Aiken et al. (2015) and Davey et al. (2015). *International Journal of Epidemiology*, 2015. doi: 10.1093/ije/dyv129.

P. W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.

S. Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

K. Imai and I. S. Kim. On the use of two-way fixed effects regression models for causal inference with panel data. *Political Analysis*, forthcoming(na):na–na, 2020.

G. Imbens and K. Kalyanaraman. Optimal Bandwidth Choice for the Regression Discontinuity Estimator. *Review of Economic Studies*, 79(3):933–959, 2012.

G. W. Imbens. Better late than nothing: Some comments on deaton (2009) and heckman and urzua (2009). *Journal of Economic Literature*, 48(2):399–423, 2010.

G. W. Imbens and J. D. Angrist. Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475, 1994.

G. W. Imbens and J. M. Wooldridge. Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, 47(1):5–86, 2009.

R. Jensen. The Digital Provide: Information (Technology), Market Performance, and Welfare in the South Indian Fisheries Sector*. *The Quarterly Journal of Economics*, 122(3):879–924, 08 2007.

R. Jensen. The (Perceived) Returns to Education and the Demand for Schooling. *The Quarterly Journal of Economics*, 125(2):515–548, 2010.

A. Kahn-Lang and K. Lang. The promise and pitfalls of differences-in-differences: Reflections on 16 and pregnant and other applications. *Journal of Business & Economic Statistics*, 0(0): 1–14, 2019.

G. King and R. Nielsen. Why propensity scores should not be used for matching. *Political Analysis*, 27(4):435–454, 2019. doi: 10.1017/pan.2019.11.

H. J. Kleven and M. Waseem. Using notches to uncover optimization frictions and structural elasticities: Theory and evidence from pakistan. *The Quarterly Journal of Economics*, 128 (2):669–723, 2013.

R. Koenker and K. F. Hallock. Quantile Regression. *Journal of Economic Perspective*, 15(4): 143–156, 2001.

C. Landais. Assessing the Welfare Effects of Unemployment Benefits Using the Regression Kink Design. *American Economic Journal: Economic Policy*, 7(4):243–78, November 2015.

E. E. Leamer. *Specification Searches – Ad Hoc Inference with Nonexperimental Data*. John Wiley & Sons, Inc., 1978.

D. S. Lee and T. Lemieux. Regression discontinuity designs in economics. *Journal of Economic Literature*, 48(2):281–355, 2010.

M.-J. Lee. *Micro-Econometrics for Policy, Program, and Treatment Effects*. Oxford University Press, 2008.

E. L. Lehmann and J. P. Romano. *Testing Statistical Hypotheses*. Springer, 2005.

J. Ludwig and D. L. Miller. Does Head Start improve children's life chances? Evidence from a regression discontinuity design. *The Quarterly Journal of Economics*, 122(1):159–208, 2000.

J. Mackinnon and M. Webb. Wild bootstrap inference for wildly different cluster sizes. *Journal of Applied Econometrics*, 32:233–254, 2017.

J. Mackinnon and M. Webb. The wild bootstrap for few (treated) clusters. *The Econometrics Journal*, 21:114–135, 11 2018. doi: 10.1111/ectj.12107.

J. McCrary. Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, 142(2):698–714, February 2008.

E. Miguel and M. Kremer. Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities. *Econometrica*, 72(1):159–217, 01 2004.

E. Miguel, C. Camerer, K. Casey, J. Cohen, K. M. Esterling, A. Gerber, R. Glennerster, D. P. Green, M. Humphreys, G. Imbens, D. Laitin, T. Madon, L. Nelson, B. A. Nosek, M. Petersen, R. Sedlmayr, J. P. Simmons, U. Simonsohn, and M. Van der Laan. Promoting Transparency in Social Science Research. *Science*, 343(6166):30–31, Jan. 2014.

G. Miller. Women's Suffrage, Political Responsiveness, and Child Survival in American History. *The Quarterly Journal of Economics*, 123(3):1287–1327, 2008.

B. Moulton. Random Group Effects and the Precision of Regression Estimates. *Journal of Econometrics*, 32(3):385–397, 1986.

R. Munroe. SIGNIFICANT (xkcd). https://xkcd.com/882/ Accessed 03 February 2017, 2010.

K. Muralidharan and N. Prakash. Cycling to School: Increasing Secondary School Enrollment for Girls in India. NBER Working Papers 19305, National Bureau of Economic Research, Inc, Aug. 2013. URL https://ideas.repec.org/p/nbr/nberwo/19305.html.

K. R. Murphy, B. Myors, and A. Wollach. *Statistical Power Analysis*. Routledge, 2014.

R. B. Newson. Frequentist q-values for multiple-test procedures. *The Stata Journal*, 10(4): 568–584, 2010.

B. A. Olken. Promises and Perils of Pre-analysis Plans. *Journal of Economic Perspectives*, 29 (3):61–80, Summer 2015.

O. Ozier. The impact of secondary schooling in Kenya: A regression discontinuity analysis. Unpublished, University of California at Berkeley, 2011.

Z. Pei, D. S. Lee, D. Card, and A. Weber. Local Polynomial Order in Regression Discontinuity Designs. *Journal of Business & Economic Statistics*, 0(0):1–9, 2021. doi: 10.1080/07350015.2021.1920961.

A. Rambachan and J. Roth. An honest approach to parallel trends, 2019.

J. P. Romano and M. Wolf. Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, 100(469):94–108, 2005a.

J. P. Romano and M. Wolf. Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4):1237–1282, 2005b.

J. P. Romano, A. M. Shaikh, and M. Wolf. Hypothesis Testing in Econometrics. *Annual Review of Economics*, 2(1):75–104, 2010.

D. Roodman. BOOTTEST: Stata module to provide fast execution of the wild bootstrap with null imposed. Statistical Software Components, Boston College Department of Economics, Dec. 2015. URL `https://ideas.repec.org/c/boc/bocode/s458121.html`.

D. Roodman, M. Ø. Nielsen, J. G. MacKinnon, and M. D. Webb. Fast and wild: Bootstrap inference in stata using boottest. *The Stata Journal*, 19(1):4–60, 2019.

P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

J. Roth. Pre-test with caution: Event-study estimates after testing for parallel trends, 2019.

A. D. Roy. SOME THOUGHTS ON THE DISTRIBUTION OF EARNINGS. *Oxford Economic Papers*, 3(2):135–146, 06 1951. doi: 10.1093/oxfordjournals.oep.a041827.

K. Schmidheiny and S. Siegloch. On event study designs and distributed-lag models: Equivalence, generalization and practical implications. IZA Discussion Papers 12079, Institute of Labor Economics (IZA), 2019.

M. Simonsen, L. Skipper, and N. Skipper. Price sensitivity of demand for prescription drugs: Exploiting a regression kink design. *Journal of Applied Econometrics*, 31(2):320–337, 2016.

L. Sun and S. Abraham. Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, 2020. ISSN 0304-4076. doi: https://doi.org/10.1016/j.jeconom.2020.09.006.

C. Taber and R. Vejlin. Estimation of a roy/search/compensating differential model of the labor market. *Econometrica*, 88(3):1031–1069, 2020. doi: https://doi.org/10.3982/ECTA14441.

M. Urquiola and E. Verhoogen. Class-size caps, sorting, and the regression-discontinuity design. *American Economic Review*, 99(1):179–215, 2009.

H. White. A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica*, 48(4):817–838, 1980.

J. M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. The MIT Press, Cambridge, Massachusetts, 2002.