Ivan A. Canay

Identification and Inference

in Modern Econometrics

| ***Econ 480-3 :: Ver. June 2, 2026***

Contents

Foreword **xv**

I Causality and Conditional Independence **1**

1 Causality and Potential Outcomes **3**
1.1 Causality, Counterfactuals, and What-if 3
1.2 Potential Outcomes 4
1.3 General Setting 6
1.4 Identification via Random Assignment 8
1.5 Estimation via Difference in Means 9
1.6 Scope of Random Assignment 11
1.7 Concluding Remarks 12
1.8 Problems 12

2 Linear Regression **15**
2.1 Interpretations of the Linear Regression Model 15
2.1.1 Interpretation 1: Linear Conditional Expectation . . . 15
2.1.2 Interpretation 2: "Best" Linear Approximation to the Conditional Expectation or "Best" Linear Predictor . . 16
2.1.3 Interpretation 3: Causal Model 17
2.2 Linear Regression and the ATE 17
2.3 Linear Regression when with Exogenous Regressors 19
2.3.1 Solving for β 19
2.3.2 Interpretation as the ATE 20
2.4 Estimating β 21
2.4.1 Ordinary Least Squares 21
2.4.2 Projection Interpretation 23
2.5 Concluding Remarks 24
2.6 Problems 24

3 More on Linear Regression **25**
3.1 Solving for Sub-vectors of β 25
3.2 Estimating Sub-Vectors of β 26
3.3 Application to Saturated in Discrete Covariates 27
3.4 Covariance Adjustment under Random Assignment 30
3.5 Concluding Remarks 35
3.6 Problems 35

4 Selection on Observables **39**
4.1 Observational Studies and Selection Bias 39
4.2 Selection on Observables 42
4.3 Estimation of the ATE 44
4.3.1 Matching 45
4.3.2 Regression 47
4.4 Concluding Remarks 50
4.5 Problems 50

5 Selection on Observables II **53**
5.1 The Role of the Propensity Score 53
5.1.1 Propensity Score Stratification 54
5.1.2 Inverse Probability Weighting 57
5.2 On the Asymptotic Efficiency of IPW Estimators 59
5.3 Multivalued Treatments 61
5.4 Scope of Selection on Observables 62
5.5 Concluding Remarks 63
5.6 Problems 63

6 Augmented IPW and Double Robustness **65**
6.1 Introduction 65
6.2 Semi-parametric Efficiency 66
6.3 Augmented IPW 67
6.4 Semiparametric Efficiency of the AIPW Estimator 69
6.5 Cross-fitting 70
6.6 Confidence Intervals 73
6.7 Concluding Remarks 73
6.8 Problems 74

II Causality and Endogeneity **77**

7 Endogeneity **79**
7.1 Endogeneity in Linear Regression 79
7.1.1 Omitted Variables 80
7.1.2 Measurement Error 81
7.1.3 Simultaneity 82
7.2 Instrumental Variables 83
7.2.1 Partition of β: solve for endogenous components . . . 86
7.3 Estimating β 86
7.3.1 The Instrumental Variables (IV) Estimator 87
7.3.2 The Two-Stage Least Squares (TSLS) Estimator . . . 88
7.4 Some examples of IVs in practice 89
7.5 Concluding Remarks 90
7.6 Problems 91

8 Properties of Two Stages Least Squares **95**
8.1 Properties of the TSLS Estimator . . . 95
8.1.1 Consistency . . . 95
8.1.2 Limiting Distribution . . . 96
8.1.3 Estimation of $\mathbb{V}$ . . . 96
8.2 Efficiency of the TSLS Estimator . . . 97
8.3 TSLS for Binary Endogenous Variables . . . 98
8.4 Empirical Applications . . . 100
8.4.1 Angrist and Evans . . . 100
8.4.2 Baum-Snow . . . 102
8.5 "Weak" Instruments . . . 104
8.6 Concluding Remarks . . . 106
8.7 Problems . . . 106

9 Heterogeneous and Endogenous Treatments **109**
9.1 A Simple Random Coefficients Model . . . 109
9.2 Wald Estimand, Heterogeneity, and LATE . . . 110
9.2.1 The Importance of Monotonicity . . . 113
9.3 An Application: Angrist and Evans revisited . . . 114
9.4 LATE's Generality . . . 116
9.4.1 Multivalued Instruments . . . 117
9.4.2 LATE with covariates . . . 118
9.5 Concluding Remarks . . . 119
9.6 Problems . . . 120

10 Marginal Treatment Effects **123**
10.1 Roy Models . . . 123
10.2 Vytlacil's Equivalence Theorem . . . 124
10.3 Marginal Treatment Effects . . . 125
10.3.1 The Normalization . . . 125
10.3.2 Definition and Interpretation . . . 126
10.4 Target Parameters as Weighted Averages of the MTE . . . 126
10.4.1 Average Treatment Effect . . . 127
10.4.2 Average Treatment Effect on the Treated . . . 127
10.4.3 Average Treatment Effect on the Untreated . . . 127
10.4.4 Local Average Treatment Effect . . . 128
10.5 Identification of the MTE . . . 128
10.5.1 Nonparametric Identification and Its Limits . . . 128
10.5.2 Parametric Identification via the MTR . . . 129
10.6 Application: Brinch, Mogstad, and Wiswall (2017) . . . 130
10.7 Concluding Remarks . . . 133
10.8 Problems . . . 133

11 Partial Identification **135**
11.1 Worst-Case Bounds . . . 135
11.1.1 Monotone Treatment Selection . . . 136
11.1.2 Monotone IV . . . 137
11.2 Sharp Bounds under IV Assumptions . . . 139
11.2.1 Roy model bounds . . . 139
11.2.2 Manski's IV bounds and their equivalence to the Roy model bounds . . . 140
11.3 Application: Bounds in the Q-Q Model . . . 141
11.4 Partial Identification in the MTE Framework . . . 142
11.4.1 Setup and target parameters . . . 143
11.4.2 Observational constraints and IV-like estimands . . . 143
11.4.3 Sharp bounds as a linear program . . . 144
11.4.4 Sharpness and the role of instrument support . . . 144
11.5 Concluding Remarks . . . 145
11.6 Problems . . . 145

III Widespread Causal Inference Designs **149**

12 Panel Data **151**
12.1 Fixed Effects . . . 152
12.1.1 First Differences . . . 152
12.1.2 Deviations from Means . . . 153
12.1.3 Asymptotic Properties . . . 154
12.2 Random Effects . . . 156
12.3 Dynamic Models . . . 158
12.4 Acknowledgement . . . 160
12.5 Problems . . . 160

13 Difference in Differences **163**
13.1 Two Groups and Two Periods . . . 163
13.1.1 Pre and post comparison . . . 165
13.1.2 Treatment and control comparison . . . 165
13.1.3 Taking both differences . . . 165
13.2 Standard Framework in DiD Models . . . 167
13.3 The Two Way Fixed Effects Estimator . . . 169
13.3.1 Basic DiD Design . . . 170
13.3.2 Staggered DiD Design . . . 172
13.4 Empirical Illustration . . . 174
13.5 Final Remarks . . . 175
13.6 Problems . . . 175

14 More Difference in Differences **179**
14.1 Event Studies in the Basic Design . . . 179
14.1.1 Application to Benzarti and Carloni (2019) . . . 181
14.2 Event Studies in Staggered Designs . . . 182
14.3 Relaxing Parallel Trends . . . 184
14.4 Synthetic Controls . . . 186
14.5 Concluding Remarks . . . 189
14.6 Problems . . . 190

15 RDD **193**
15.1 Sharp RDD: Identification . . . 193
15.2 Estimation via Local Linear Regression . . . 196
15.2.1 Bandwidth Choice . . . 197
15.3 Inference . . . 198
15.3.1 Conventional Inference and Undersmoothing . . . 199
15.3.2 Standard Bias Correction . . . 199
15.3.3 Robust Bias Correction . . . 200
15.4 Empirical Example . . . 201
15.5 Validation and Extensions . . . 203
15.6 Concluding Remarks . . . 203
15.7 Problems . . . 203

IV Inference via Resampling **205**

16 Randomization Tests **207**
16.1 The Classical Fisher Framework . . . 207
16.1.1 The FRT . . . 208
16.1.2 The choice of test statistic . . . 210
16.2 Design-based vs Sampling-based Uncertainty . . . 211
16.3 Randomization Tests . . . 213
16.3.1 Motivating example: sign changes . . . 213
16.3.2 The main result . . . 214
16.3.3 Permutation tests for treatment effects . . . 217
16.4 Asymptotic validity of permutation tests . . . 219
16.5 Concluding Remarks . . . 220
16.6 Problems . . . 220

17 The Bootstrap **223**
17.1 Confidence Sets . . . 223
17.1.1 Pivots and Asymptotic Pivots . . . 225
17.1.2 Asymptotic Approximations . . . 226
17.2 The Bootstrap . . . 227
17.2.1 The Nonparametric Mean . . . 228
17.2.2 Asymptotic Refinements . . . 232
17.2.3 Implementation of the Bootstrap . . . 232
17.3 Concluding Remarks . . . 233

17.4 Problems . . . 234

18 Inference with Clustered Data **237**
18.1 Setup and Notation . . . 237
18.2 Law of Large Numbers . . . 238
18.3 Central Limit Theorem . . . 239
18.3.1 Cluster Covariance Estimation . . . 240
18.4 Inference for Linear Regression: CCE Approach . . . 241
18.4.1 Cluster Covariance Estimator . . . 242
18.4.2 Inference . . . 242
18.4.3 The Problem with Few Clusters . . . 243
18.5 The Wild Bootstrap . . . 244
18.5.1 The test . . . 244
18.5.2 Asymptotic validity of the test . . . 246
18.6 When and At What Level to Cluster? . . . 247
18.7 Supplement: rates of convergence . . . 248
18.8 Concluding Remarks . . . 251
18.9 Problems . . . 251

V Topics not covered in class **253**

A Properties of LS **255**
A.1 Properties of LS . . . 255
A.1.1 Bias . . . 255
A.1.2 Gauss-Markov Theorem . . . 256
A.1.3 Consistency . . . 257
A.1.4 Limiting Distribution . . . 257
A.2 Estimation of $\mathbb{V}$ . . . 258
A.3 Improving finite sample performance: HC2 & HC3 . . . 259
A.4 Consistency of HC standard errors . . . 261
A.5 Measures of Fit . . . 263
A.6 Concluding Remarks . . . 264
A.7 Problems . . . 264

B Basic Inference **267**
B.1 Inference . . . 267
B.1.1 Background . . . 267
B.1.2 Tests of A Single Linear Restriction . . . 268
B.1.3 Tests of Multiple Linear Restrictions . . . 269
B.1.4 Tests of Nonlinear Restrictions . . . 270

C More RDD **273**
C.1 RD Plots . . . 273
C.1.1 Choosing the Location of Bins . . . 275
C.1.2 Choosing the Number of Bins . . . 275
C.2 Validation and Falsification of the RD Design . . . 276

C.2.1 Density of Running Variable 277
C.2.2 Predetermined Covariates and Placebo Outcomes . . . 278
C.2.3 Placebo Cutoffs 279
C.2.4 Sensitivity to Observations near the Cutoff 280
C.3 The Fuzzy RD Design 280
C.4 Concluding Remarks 282
C.5 Problems 283

Foreword

These lecture notes were prepared for Econ 480-3 at Northwestern University, the third quarter in the first-year sequence in Econometrics. The class assumes that students have taken Econ 480-1, covering the fundamentals of asymptotic theory and M-estimation, and Econ 480-2, which includes a wide range of estimation tools for prediction problems, including parametric, semiparametric, and non-parametric approaches (including modern machine learning techniques). In this context, the focus of this class is not on technicalities but rather on developing a solid understanding of the situations where we can assign causal interpretations to commonly used parameters that applied researchers seek to estimate.

There are numerous resources available online and in print that cover many of the topics in these notes in significant more detail and depth, and each chapter strives to clearly articulate and highlight the references that influenced the ideas presented. Why write these notes then? The primary purpose is to collect the main ideas I aim to cover in the class within a unified framework and notation. Personally, I do not believe that these notes would be particularly useful outside the context of Econ 480-3 and are not intended as a comprehensive treatment of any of the topics covered. You should feel free to use them at your own risk, but I highly encourage readers to explore and consult the additional resources referenced in each chapter.

I would like to express my gratitude to several friends and colleagues who made it easier for me to write these notes. First and foremost, I want to thank Stefan Wager and Peng Ding for their excellent notes on many of the topics covered here, which they have generously made publicly available. I also want to extend my gratitude to those who were kind enough to share their own notes with me, even when such notes are not publicly available. These individuals include Alex Torgovitsky for his outstanding notes on IV, Clément de Chaisemartin for sharing his work in progress on Difference in Differences, Matias Cattaneo for his teaching material on RDD, Matt Masten for his notes on partial identification, and Chuck Manski for sharing his notes on the previous iteration of Econ 480-1 that is now part of second year classes. Special thanks go out to my friend and long-time collaborator, Azeem Shaikh, whose influence on my way of thinking about econometrics makes it challenging for me to distinguish how much of what is written in these notes is novel and how much is derived from our conversations, scribbles, and the notes he has shared with me over our 17+ years of collaboration. Finally, I want to express my sincere appreciation to Shuyan Huang, whose invaluable contributions as my research assistant played a key role in developing and revising these notes.

Part I

Causality and Conditional Independence

1

Causality and Potential Outcomes

If you are graduate student in an economics program reading these notes you are already innately familiar with the concept of causality and the difference between association/correlation and causation. Since you already understand the definition of causal effect and the difference between association and causation, do not expect to gain deep conceptual insights from this lecture. Rather, the purpose of this lecture is to introduce mathematical notation that formalizes the causal intuition that you already possess. Our goal is to make sure that you can match your causal intuition with the mathematical notation introduced here and, moving down the road help you to understand *when* a certain parameter may be interpreted causally or not. This notation is necessary to precisely define causal concepts, and will be used throughout the class.

1.1 Causality, Counterfactuals, and What-if

In this class we will use the so-called potential outcomes framework (Neyman, 1923; Rubin, 1974) to mathematically represent counterfactual states of the world. In this framework, an experiment, or at least a thought experiment, has an intervention, a manipulation, or a treatment, and we are interested in its effect on an outcome or multiple outcomes. Let's start with some examples of counterfactual questions that appear in different areas of econ and even outside econ:

What happens if a job training program is expanded?	Labor
What would happen to prices/welfare if two firms merged?	IO
What would a different monetary policy do to real output?	Macro
What effect would this medication have on heart disease?	Biostatistics
What will happen to global temps if emissions decrease?	Climatology

Suppose two firms, AA and UA, merge in 2022. Let p^{obs} denote the price one year after the merger, and let p^{cf} denote the price that would have prevailed one year later had the firms not merged (holding fixed the relevant

environment). If p^{obs} exceeds the pre-merger price while p^{cf} would have remained unchanged, then the merger has a positive causal effect on price over this horizon.

Now consider a different merger, between firms C and D, also in 2022. Again suppose the post-merger price rises relative to the previous year. If, however, the no-merger counterfactual p^{cf} would have risen by exactly the same amount, then the merger has no causal effect on price (over this horizon).

These vignettes capture the basic logic of causal effects: compare the realized outcome under an action to the outcome that would have occurred had the action been withheld. The action A may be a policy, intervention, exposure, or treatment; the causal effect is defined as the contrast between these two (potential) outcomes.

1.2 Potential Outcomes

Probably the easiest way to think about causal relationships is in terms of potential outcomes. As a simple illustration, consider a randomized controlled experiment where individuals are randomly assigned to a treatment (a drug) that is intended to improve their health status. Let Y denote the observed health status and $A \in \{0, 1\}$ denote whether the individual takes the drug or not. The causal relationship between A and Y can be described using the so-called *potential outcomes*:

$Y(0)$	potential outcome in the absence of treatment
$Y(1)$	potential outcome in the presence of treatment

In other words, we imagine two potential health status variables $(Y(0), Y(1))$ where $Y(0)$ is the value of the outcome that would have been observed if (possibly counter-to-fact) A were 0; and $Y(1)$ is the value of the outcome that would have been observed if (possibly counter-to-fact) A were 1.

The difference

$$\Delta := Y(1) - Y(0) \tag{1.1}$$

is called the *treatment effect*, and the quantity $E[Y(1) - Y(0)]$ is usually referred to as the *average treatment effect*. Our discussion for now will focus on population parameters and so we will not talk about the availability of a random sample or index random variables by subjects, units, or individuals; i.e., by i. If we were to do so, we would write $Y_i(0)$ and $Y_i(1)$ for the potential outcomes of the i-th individual and the treatment effect in (1.1) would be indexed by Δ_i. This would indicate that the treatment effect is unit specific and potentially heterogeneous across individuals. Neyman (1923) first used this notation and, as simple and intuitive as it appears to be, it has some hidden assumptions. Rubin (1980) clarified the implicit assumptions and called them no-interference and consistency.

The first assumption, no-interference, states that unit i's potential outcomes do not depend on other units' treatments. This is often a reasonable assumption in medicine (i.e., that the treatment prescribed to patient 1 doesn't affect patient 2), but may be less appropriate in social sciences where network effects may arise. For now, however, we will ignore network effects and assume no-interference.

The second assumption, consistency, states that there are no other versions of the treatment. Equivalently, it requires that the treatment level be well defined, or have no ambiguity at least for the outcome of interest. For instance, when studying the effect of cigarette smoking on lung cancer, the type of cigarettes may matter; when studying the effect of college education on income, the type and major of college education may matter. Consistency assumes away such differences.

The two assumptions combined imply that the observed outcome Y_i for every unit i is equal to the potential outcome $Y_i(A_i)$ for the treatment level $A_i \in \{0, 1\}$ that was actually assigned to unit i,

$$Y_i = Y_i(A_i) = Y_i(1)A_i + Y_i(0)(1 - A_i) \ .$$

We can re-write this equation as

$$Y_i = Y_i(0) + \Delta_i A_i \ ,$$

which shows that the observed outcome is just the baseline potential outcome $Y(0)$ plus the treatment effect times the treatment level. Importantly, all of these objects, $Y(0)$, A and Δ, are random. In particular, since Δ is random, it has a distribution and we can talk about its mean, variance, and other properties. In other words, the treatment effects can be summarized in many ways and in turn lead to many possible *parameters of interest.*

Remark 1.1 There are two related but distinct choices in causal analysis. One is whether the target is a superpopulation quantity or a finite-population quantity for the realized sample. The other is whether randomness comes from sampling/modeling assumptions or from the treatment assignment itself. In these notes we usually begin with a superpopulation formulation, so $(Y(0), Y(1), A, \Delta)$ are random variables, and later we will indicate when we switch to statements that condition on the realized sample or on the assignment mechanism. ■

Example 1.1 (Program Evaluation) Suppose that $A \in \{0, 1\}$ indicates participation in a job training program and that Y is a scalar labor market outcome such as earnings. Under the potential outcomes framework, each unit has $(Y(0), Y(1))$, where $Y(a)$ is the outcome that would be observed under treatment status $A = a$. We observe $Y = Y(A)$, so if $A = 1$ we observe $Y(1)$ (but not $Y(0)$), and if $A = 0$ we observe $Y(0)$ (but not $Y(1)$). Examples of causal questions include:

- $E[Y(1)]$: What would average earnings be if everyone were trained?
- $E[Y(1) - Y(0)]$: What is the average effect of the program?
- $E[Y(1) - Y(0) \mid A = 1]$: What is the average effect for those who are trained?

■

Our goal in this class will be to credibly identify and estimate features of the distribution of the treatment effect Δ. We will devote particular attention to the average treatment effect (ATE) defined as

$$\theta := E[\Delta] = E[Y(1) - Y(0)] \ ,$$

due to its prevalence in empirical work. However, as we will discuss throughout the course, the ATE is not always the most interesting parameter to consider and may not even be policy relevant.

The main barrier to credibly identify features of the distribution of the treatment effect Δ, such as the ATE, is that only one treatment can be assigned to a given individual, and so only one of $Y(0)$ and $Y(1)$ can ever be observed. In other words, the treatment effect Δ is *never* observed and so we need to find ways to deal with this missing data problem. What we observe is the outcome Y for a given treatment level A, and so the problem of identification is to be able to characterize θ as a function of the observed data (Y, A).

1.3 General Setting

The potential outcomes framework is a way to formalize the notion of counterfactuals. While for the vast majority of our exposition we will focus on the case where the treatment variable A is binary, it is useful to think about the case where A can take on more than two values. For instance, in the case of the effect of education on income, A could be the number of years of education. In the general case we will let $\mathcal{A}$ denote the support of A, where $\mathcal{A}$ is a mutually exclusive and exhaustive set of states (the treatments). To keep notation simple, throughout this chapter we will focus on the case where A has a finite or countable support $\mathcal{A}$ (so that sums such as $\sum_{a \in \mathcal{A}}$ are well defined). We return to continuous treatments later in the course. Examples include training/no training $\mathcal{A} = \{0, 1\}$ and years of education $\mathcal{A} = \{0, 1, \ldots, 25\}$.

$$Y(a) = \text{ the outcome that would have been observed if } A = a \ .$$

The treatment effect in this more general case could be defined relative to any baseline $a_0 \in \mathcal{A}$ as $\Delta(a) = Y(a) - Y(a_0)$, or between two different values of

a as $\Delta(a, a') = Y(a) - Y(a')$. In the general setting, the observed outcome Y satisfies $Y = Y(A)$, and in the case where A is discrete we can write

$$Y = \sum_{a \in \mathcal{A}} Y(a) I\{A = a\} = Y(A) \ ,$$

where $I\{A = a\}$ is an indicator function that takes the value 1 if $A = a$ and 0 otherwise. Note that $Y = Y(A)$ is observed, but $Y(a)$ for $a \neq A$ are always **unobserved**. We are interested in features of the **counterfactuals** $Y(a)$ for $a \neq A$, as these counterfactuals in turn allow us to define parameters of interest such as the average treatment effect.

Before we move on to discuss identification of the ATE in perhaps the simplest possible setting, we should mention that potential outcomes are not the only way to formalize the notion of counterfactuals. There are at least two other ways that receive a lot of attention across disciplines: structural models (or latent variable models) and directed acyclic graphs (DAGs). Latent variable models generally refer to models where the outcome is a function of a latent variable and some other variables. For instance, in the case of the effect of education A on income Y, the latent variable could be IQ (denoted by U), and be related to the outcome by the relationship

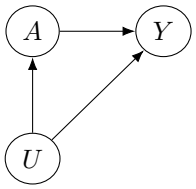
$$Y = g(A, U) \ ,$$

for some well defined function $g(\cdot)$. A **causal** interpretation of this model is implicitly saying:

$$Y(a) = g(a, U) \text{ for every } a \in \mathcal{A}$$

and it could impose assumptions depending on what g and U are (and, more importantly, how U is assumed to be related to A).

DAGs are a graphical representation of a set of assumptions about the causal structure of the data generating process. For example, in the case of the effect of education A on income Y, we could have the following DAG



where an arrow represents a direct causal effect. For instance, the arrow from A to Y encodes the assumption that education has a direct causal effect on income. The arrows from U to A and Y encode the assumption that unobserved ability affects both education and income, which is one way to represent endogeneity/selection into education. Importantly, the absence of an arrow is a statement about *direct causal* effects within the posited model; it does not, by itself, imply statistical independence (and it does not rule out dependence

induced by omitted variables that are not shown in the graph). For many, a DAG is a compact way to record causal assumptions.

Some are dogmatic about the use of, say, potential outcomes versus latent variable models. Often follows some field-specific social norms, like in labor versus IO. We will not discuss these alternative approaches in detail in this class, but you should be aware of them and keep in mind that this is just *notation* - and that you can use the above to translate.

1.4 Identification via Random Assignment

The main element in our search for identification is the assumption we make on the treatment (or intervention) variable A. Random assignment is the assumption that the treatment variable A is independent of the potential outcomes $Y(a)$ for all $a \in \mathcal{A}$. In other words, random assignment is the assumption that

$$\{Y(a) : a \in \mathcal{A}\} \perp\!\!\!\perp A \ . \tag{1.2}$$

Under random assignment, the distribution of $Y(a)$ is identified,

$$F_a(y) := P\{Y(a) \leq y\} = P\{Y(a) \leq y \mid A = a\} = P\{Y \leq y \mid A = a\} \ , \tag{1.3}$$

where the second equality is due to random assignment. The intuition of this step is that under random assignment, conditioning on treatment does not change the distribution of potential outcomes. In other words, there is nothing systematically different about the treatment and control groups. Identification follows from the last equality since the distribution of Y given $A = a$ is identified from the data. It follows from this result that any parameter that is a function of $\{F_a : a \in \mathcal{A}\}$ is also identified. Some common parameters of interest when A is binary are:

Average treatment effect (ATE)	$E[Y(1) - Y(0)]$
Average treatment on the treated (ATT)	$E[Y(1) - Y(0) \mid A = 1]$
Average treatment on the untreated (ATU)	$E[Y(1) - Y(0) \mid A = 0]$
Quantile treatment effect (QTE)	$Q_{Y(1)}(t) - Q_{Y(0)}(t)$
QTE on treated/untreated (QTT/QTU)	defined analogously

Under random assignment, all of these parameters are identified and $ATE = ATT = ATU$, and $QTE = QTT = QTU$.

To see more directly how the ATE is identified by the distribution of the observed data (Y, A) under random assignment, note that

$$\begin{aligned} \theta &:= E[Y(1) - Y(0)] \\ &= E[Y(1) \mid A = 1] - E[Y(0) \mid A = 0] \\ &= E[Y \mid A = 1] - E[Y \mid A = 0] \ , \end{aligned} \tag{1.4}$$

where the second equality follows from random assignment and the third equality follows from $Y = Y(1)A + (1 - A)Y(0)$.

Even under the assumption in (1.2), the joint distribution of the potential outcomes $Y(0)$ and $Y(1)$ is not identified, i.e.,

$$P\{Y(1) \leq y_1, Y(0) \leq y_0\} \ .$$

This is because we never observe both potential outcomes for the same unit and is known as the *fundamental problem of causal inference*. Importantly, most features of $\Delta = Y(1) - Y(0)$ are not (point) identified. For instance, we cannot identify the proportion of individuals who are hurt by the treatment, $P\{Y(1) \leq Y(0)\}$, nor the quantiles of Δ. Note that this last observation highlights the differences between the quantile treatment effects, QTE, versus the quantiles of the treatment effect (the quantiles of Δ).

1.5 Estimation via Difference in Means

Consider the setting from a randomized controlled experiment (or trial, i.e., RCT) where we have a binary treatment variable A and an outcome variable Y. We assume that we have a random sample of size n from the distribution of (Y, A), which we denote by P, and that A is exogenous in the sense of (1.2). We also assume that we have treated and control units, i.e., $n_1 > 0$ and $n_0 > 0$ a.s.

In the previous section we characterized the ATE parameter θ as a function of (Y, A) in (1.4). Given this representation, the natural estimator of the ATE is the difference in means estimator given by

$$\hat{\theta}_n := \frac{1}{n_1} \sum_{i=1}^{n} Y_i A_i - \frac{1}{n_0} \sum_{i=1}^{n} Y_i (1 - A_i) = \frac{1}{n_1} \sum_{i \in \mathcal{I}_1} Y_i - \frac{1}{n_0} \sum_{i \in \mathcal{I}_0} Y_i \ , \qquad (1.5)$$

where we used the notation

$$\mathcal{I}_a := \{i \in \{1, \ldots, n\} : A_i = a\}$$

to denote the set of units with treatment status a and $n_a := |\mathcal{I}_a|$ to denote the number of units with treatment status a. That is, the estimator in (1.5) is the difference in sample means of the outcome variable between the treatment and control groups.

Under the random assignment assumption in (1.2), the estimator in (1.5) is unbiased for θ; that is

$$E[\hat{\theta}_n] = \theta \ .$$

To see this, first note that

$$\begin{aligned} E[\hat{\theta}_n] &\overset{(1)}{=} E\left[E[\hat{\theta}_n \mid A_1, ..., A_n]\right] \\ &\overset{(2)}{=} E\left[E\left[\frac{1}{n_1}\sum_{i=1}^{n} Y_i A_i \mid A_1, \ldots, A_n\right] - E\left[\frac{1}{n_0}\sum_{i=1}^{n} Y_i(1-A_i) \mid A_1, \ldots, A_n\right] \right] , \end{aligned}$$

where $\overset{(1)}{=}$ follows from the law of iterated expectations (LIE), and $\overset{(2)}{=}$ follows from the linearity of the expectation. Then, taking the conditional expectation of the treatment group average (i.e. the first term above) we get:

$$\begin{aligned} E\left[\frac{1}{n_1}\sum_{i=1}^{n} Y_i A_i \mid A_1, \ldots, A_n\right] &\overset{(3)}{=} \frac{1}{n_1}\sum_{i=1}^{n} E[A_i Y_i \mid A_1, \ldots, A_n] \\ &\overset{(4)}{=} \frac{1}{n_1}\sum_{i=1}^{n} E[A_i Y_i(1) \mid A_1, \ldots, A_n] \\ &\overset{(5)}{=} \frac{1}{n_1}\sum_{i=1}^{n} A_i E[Y_i(1) \mid A_1, \ldots, A_n] \\ &\overset{(6)}{=} \frac{1}{n_1}\sum_{i=1}^{n} A_i E[Y_i(1) \mid A_i] \\ &\overset{(7)}{=} \frac{1}{n_1}\sum_{i=1}^{n} A_i E[Y_i(1)] \\ &\overset{(8)}{=} E[Y(1)] , \end{aligned}$$

where $\overset{(3)}{=}$ follows from the linearity of the expectation; $\overset{(4)}{=}$ follows from $A_i Y_i = A_i Y_i(1)$; $\overset{(5)}{=}$ uses that A_i is non-random conditional on $(A_1, \ldots, A_n)$; $\overset{(6)}{=}$ follows from i.i.d. sampling across units; $\overset{(7)}{=}$ uses random assignment; and $\overset{(8)}{=}$ uses that $\sum_{i=1}^{n} A_i = n_1$. Similar arguments show that

$$E\left[\frac{1}{n_0}\sum_{i=1}^{n} Y_i(1-A_i) \mid A_1, \ldots, A_n\right] = E[Y(0)] ,$$

which implies that $E[\hat{\theta}_n] = E[Y(1)] - E[Y(0)] = \theta$ by the law of iterated expectations (LIE).

Later in this class we will show that this estimator is also consistent and asymptotically normal under these assumptions.

Example 1.2 Bertrand and Mullainathan (2004) Bertrand and Mullainathan [2004] conducted a randomized experiment on resumes to study the effect of perceived race on callbacks for interviews. They randomly assigned African-American - or White - sounding names on fictitious resumes to help-wanted

ads in Boston and Chicago newspapers. The following two-by-two table summarizes perceived race and callback:

	callback	no callback
African-American	157	2278
White	235	2200

We can compare the probabilities of being called back among African-American - and White - sounding names:

$$\frac{157}{2435} - \frac{235}{2435} = 6.45\% - 9.65\% = -3.20\% < 0 \ .$$

That is, white names received more callbacks. In this case we could do a Fisher's exact test (later on in class) to learn that this difference is statistically significant with a p-value smaller than 0.001. In Bertrand and Mullainathan (2004)'s experiment, the treatment is the "perceived race" (as opposed to race) which can be manipulated by experimenters and so tackles a well-defined causal question. ■

1.6 Scope of Random Assignment

One may wonder when is random assignment a good assumption. The answer is that this is a reasonable assumption only in contexts where the experimenter has control over A (as in Example 1.2) or, in other words, settings where agents have no control over A. It is much less likely to hold in settings where units choose A, since they typically choose A using information about $\{Y(a) : a \in \mathcal{A}\}$. For example, random assignment is rarely compelling with observational data where agents choose A to maximize some criterion function (utility, profits, etc.). In those cases, we expect *selection*.

We say there is selection into the treatment state A if

$$Y(a) \mid A = a \text{ is distributed } \textbf{differently} \text{ from } Y(a) \mid A = a' \text{ for } a \neq a' \ .$$

This is expected to occur if agents choose A with knowledge of $\{Y(a) : a \in \mathcal{A}\}$. For example, agents who choose to join a job training program might do so because of a low value of $Y(0)$. Alternatively, they might choose $A = 0$ because of a high value of $Y(0)$. In either case, we expect $Y(0)$ to be distributed differently across the two groups and this in turn leads to the so-called *selection bias*.

The difference in means estimator in (1.5) converges to the ATE only if there is no selection into the treatment state. In general, it always converges to the characterization we derived in (1.4),

$$E[Y \mid A = 1] - E[Y \mid A = 0] \ .$$

We can then decompose the contrast into a causal effect and selection bias:

$$E[Y \mid A=1] - E[Y \mid A=0] = E[Y(1) \mid A=1] - E[Y(0) \mid A=1] \\ + E[Y(0) \mid A=1] - E[Y(0) \mid A=0] \ .$$

The first difference on the right-hand side is the causal effect for those who were treated (the ATT), while the second difference is the selection bias term that captures how the treated would have been different anyway. These effects could cancel out if the ATT is $(+)$ while the selection bias is $(-)$, but they could also go in the same direction and exacerbate each other.

1.7 Concluding Remarks

The contents of this chapter are based on the notes by Peng Ding, Ding [2023], the notes by Stefan Wager, Wager [2020], and notes shared by Alex Torgovitsky. The book Angrist and Pischke [2008] is another reference that includes discussion on many of the topics we covered.

1.8 Problems

Problem 1.1 *Show that under random assignment we obtain*

$$ATE = ATT = ATU \ .$$

Problem 1.2 *In this problem, we look into more details of Bertrand and Mullainathan (2004)'s study on the effect of perceived race on callbacks for interviews. They first generated a pool of resumes for the fictitious job applicants. The resumes are classified into two categories: high and low quality. They also generated a pool of names for the fictitious job applicants. The names are classified into four categories: African-American male, African-American female, white male, and white female.*

1. *Suppose for each resume randomly drawn from the pool of resumes, a name is randomly drawn from the pool of names to generate a fictitious job applicant. In this case, is the random assignment (of perceived race) assumption valid for each of the following subgroups? Justify your answer.*

   (a) *high quality resume;*

   (b) *low quality resume;*

   (c) *female;*

(d) male.

2. *For the four subgroups in part 1, conduct the same analysis as in Section 1.5. That is, for each subgroup, calculate the difference in means estimator for the effect of perceived race on the probability of callback. Report the estimates in a table and explain your findings.*

3. *If multiple fictitious resumes are sent by the experimenter to the same company, which assumption in Section 1.2 is questionable?*

Problem 1.3 *Suppose* $(Y(1), Y(0))$ *are the potential outcomes of patients' health with and without some treatment, and that they are generated from*

$$Y(0) \sim \mathrm{N}(0,1), \quad \tau = -0.5 + Y(0), \quad Y(1) = Y(0) + \tau.$$

In the following two scenarios, calculate the difference in means $E(Y \mid A = 1) - E(Y \mid A = 0)$*. Explain how selection leads to the difference between the two numbers using the decomposition in Section 1.6.*

1. *A clueless doctor does not know any information about the individual causal effects and assigns the treatment to patients by flipping a fair coin.*

2. *A perfect doctor knows the individual causal effects and assigns the treatment by* $A = 1(\tau \geq 0)$*.*

Hint: The mean of a truncated Normal random variable equals

$$E(X \mid a < X < b) = \mu - \sigma \frac{\phi\left(\frac{b-\mu}{\sigma}\right) - \phi\left(\frac{a-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)},$$

where $X \sim \mathrm{N}\left(\mu, \sigma^2\right)$*, and* $\phi(\cdot)$ *and* $\Phi(\cdot)$ *are the probability density and cumulative distribution functions of a standard Normal random variable.*

Problem 1.4 *The median treatment effect*

$$\delta_1 = \text{median}\left\{Y(1)\right\} - \text{median}\left\{Y(0)\right\} \ .$$

is, in general, different from the median of the individual treatment effect

$$\delta_2 = \text{median}\left\{Y(1) - Y(0)\right\} \ .$$

1. *Give numerical examples which have* $\delta_1 = \delta_2$*,* $\delta_1 > \delta_2$*, and* $\delta_1 < \delta_2$*.*

2. *Which estimand makes more sense,* δ_1 *or* δ_2*? Why? Use examples to justify your conclusion. If you feel that both can make sense in different applications, you can also give examples to justify both estimands.*

Problem 1.5 *Consider the difference in means estimator $\hat{\theta}_n$. Assume that the treatment assignment mechanism is via what is known as complete randomization: that is, you split the n units in the experiments in two random halves and then assign one of them to treatment and the other one to control. In this case, $n_1 = n_0 = n/2$ (assume n is even for simplicity) and $E[A] = 1/2$. Show that $\hat{\theta}_n$ is unbiased for θ.*

Problem 1.6 *Let $A \in \{0, 1\}$ be a randomly assigned treatment and let $Y \in \{0, 1\}$ be a binary outcome. Assume the maintained potential-outcomes framework, and suppose random assignment holds:*

$$\{Y(0), Y(1)\} \perp\!\!\!\perp A \ .$$

Then the data identify the marginal distributions of $Y(0)$ and $Y(1)$. Let

$$P\{Y(1) = 1\} = p_1, \qquad P\{Y(0) = 1\} = p_0,$$

where $0 < p_0, p_1 < 1$.

1. *Construct two different joint distributions for $(Y(0), Y(1))$ that are both consistent with the same marginals (p_0, p_1).*
2. *For each joint distribution, compute the "harmed" fraction $P\{Y(1) < Y(0)\}$.*
3. *Conclude that $P\{Y(1) < Y(0)\}$ is not identified from the observed distribution of (Y, A), even under random assignment.*

Bibliography

J. D. Angrist and J.-S. Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2008.

M. Bertrand and S. Mullainathan. Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American Economic Review*, 94(4):991–1013, September 2004. doi: 10.1257/0002828042002561. URL https://www.aeaweb.org/articles?id=10.1257/0002828042002561.

P. Ding. A first course in causal inference. arXiv:2305.18793, 2023.

S. Wager. Causal inference. Standford University, 2020.

2

Linear Regression

In this previous lecture we introduced the notion of counterfactuals, potential outcomes, and causal parameters like the average causal effect (ATE). We also discuss how we could identify the ATE under random assignment, leading to the difference in means estimator. In this lecture we will introduce the linear regression model and discuss how it can be used to estimate causal parameters. We will also discuss the assumptions that are necessary to interpret the parameters in the linear regression model causally and how the ATE can be estimated using the linear regression model under random assignment.

2.1 Interpretations of the Linear Regression Model

Let (Y, X, U) be a random vector where Y and U take values in $\mathbf{R}$ and X takes values in $\mathbf{R}^{k+1}$. Assume further that the first component of X is a constant equal to one, i.e., $X = (X_0, X_1, \ldots, X_k)'$ with $X_0 = 1$. Let $\beta = (\beta_0, \beta_1, \ldots, \beta_k)' \in \mathbf{R}^{k+1}$ be such that

$$Y = X'\beta + U \ .$$

The parameter β_0 is sometimes referred to as the *intercept parameter* and the remaining β_j parameters are sometimes referred to as the *slope parameters*. There are several ways to interpret β depending on the assumptions imposed on (Y, X, U). We will study three such ways.

2.1.1 Interpretation 1: Linear Conditional Expectation

Suppose $E[Y|X] = X'\beta$ and define $U = Y - E[Y|X]$. (Note that we've implicitly assumed $E[|Y|] < \infty$, so $E[Y|X]$ exists.) This implies that $E[U|X] = 0$ and therefore that $E[U] = 0$. Moreover, $E[XU] = 0$, so $\text{Cov}[X, U] = 0$. In this case, β is just a convenient way of summarizing a feature of the joint distribution of Y and X, namely, the conditional expectation. It is tempting to interpret the coefficient β_j for $1 \leq j \leq k$ as the *ceteris paribus* (i.e., holding X_{-j} and U constant) effect of a one unit change in X_j on Y, but this is incorrect. Indeed, more generally, it is not appropriate to think of differences in (or derivatives of) conditional expectations causally. After all, Y could be

an indicator for rain and X could be an indicator for carrying an umbrella. In this case, it may be the case that $E[Y|X]$ is increasing in X, but one would not want to think of carrying an umbrella as causing rain. What is missing is a model of how Y is determined as a function of X (and possibly other unobserved variables).

2.1.2 Interpretation 2: "Best" Linear Approximation to the Conditional Expectation or "Best" Linear Predictor

In general, one would not expect the conditional expectation to be linear. Suppose $E[Y^2] < \infty$ and $E[XX'] < \infty$ (equivalently, that $E[X_j^2] < \infty$ for $1 \leq j \leq k$). Under these assumptions, one may consider what is the "best" linear approximation (i.e., function of the form $X'b$ for some choice of $b \in \mathbf{R}^{k+1}$) to the conditional expectation. To this end, consider the minimization problem

$$\min_{b \in \mathbf{R}^{k+1}} E[(E[Y|X] - X'b)^2] \ .$$

Denote by β a solution to this minimization problem. In this case, β is simply a convenient way of summarizing another feature of the joint distribution of Y and X, namely, the "best" linear approximation to the conditional expectation. For the same reasons as before, it is not correct to interpret the coefficient β_j for $1 \leq j \leq k$ as the *ceteris paribus* effect of a one unit change in X_j on Y.

Let $V = E[Y|X] - Y$, so $E[XV] = 0$. Note that

$$\begin{aligned} E[(E[Y|X] - X'b)^2] &= E[(E[Y|X] - Y + Y - X'b)^2] \\ &= E[(V + Y - X'b)^2] \\ &= E[V^2 + 2V(Y - X'b) + (Y - X'b)^2] \\ &= E[V^2] + 2E[VY] - 2E[VX']b + E[(Y - X'b)^2] \\ &= \text{constant} + E[(Y - X'b)^2] \ . \end{aligned}$$

Thus, β also solves

$$\min_{b \in \mathbf{R}^{k+1}} E[(Y - X'b)^2] \ .$$

In this sense, β is also a convenient way of summarizing the "best" linear predictor of Y given X. Again, it is tempting to interpret β_j for $1 \leq j \leq k$ causally, but this is not correct.

Consider the second minimization problem. Note $E[(Y - X'b)^2]$ is convex (as a function of b) and

$$D_b E[(Y - X'b)^2] = E[-2X(Y - X'b)] \ .$$

Hence, β must satisfy

$$E[X(Y - X'\beta)] = 0 \ .$$

If we define $U = Y - X'\beta$, then we may rewrite this equation as

$$E[XU] = 0 \ .$$

2.1.3 Interpretation 3: Causal Model

Suppose $Y = g(X, U)$, where X are the observed determinants of Y and U are the unobserved determinants of Y. Such a relationship is a model of how Y is determined and may come from physics, economics, etc. The effect of X_j on Y holding X_{-j} and U constant (i.e., *ceteris paribus*) is determined by g. If g is differentiable, then it is given by $D_{X_j} g(X, U)$. If we assume further that

$$g(X, U) = X'\beta + U \ ,$$

then the *ceteris paribus* effect of X_j on Y is simply β_j. We may normalize U so that $E[U] = 0$ (by replacing U with $U - E[U]$ and β_0 with $\beta_0 + E[U]$ if this is not the case). On the other hand, $E[U|X]$, $E[U|X_j]$ and $E[UX_j]$ for $1 \leq j \leq k$ may or may not equal zero. These are now statements about the relationship between the observed and unobserved determinants of Y.

It is probably fair to say that the causal interpretation of β as described above is in disuse. The main reason not only lies in the difficulty of justifying in applications that $g(\cdot)$ is linear in X, but also in the fact that the model implicitly assumes that the effect of X on Y is *homogenous* across individuals - in other words, the model assumes that every single agent responds to changes in X in the same way. However, it is also fair to say that least squares is prevalent in applied work and often interpreted as capturing *some type* of causal effect. We will explore this interpretation in the next section, after we have a characterization of β as a function of the distribution of (Y, X), and show that β can sometimes be expresses as a *weighted average* of *heterogeneous* causal effects. Interpreting β in this way is delicate and heavily depends on the specifics of the application.

2.2 Linear Regression and the ATE

As we discussed in the previous lecture, perhaps the easiest way to think about causal relationships is in terms of potential outcomes. As a simple illustration, consider a randomized controlled experiment where individuals are randomly assigned to a treatment (a drug) that is intended to improve their health status. Let Y denote the observed health status and $A \in \{0, 1\}$ denote whether the individual takes the drug or not. The causal relationship between A and Y can be described using the so-called *potential outcomes*:

$Y(0)$	potential outcome in the absence of treatment
$Y(1)$	potential outcome in the presence of treatment

.

In other words, we imagine two potential health status variables $(Y(0), Y(1))$ where $Y(0)$ is the value of the outcome that would have been observed if

(possibly counter-to-fact) A were 0; and $Y(1)$ is the value of the outcome that would have been observed if (possibly counter-to-fact) A were 1.

The difference $Y(1) - Y(0)$ is called the *treatment effect*, and the quantity $E[Y(1) - Y(0)]$ is usually referred to as the *average treatment effect*. Using this notation, we may rewrite the observed outcome as

$$\begin{aligned} Y &= AY(1) + (1 - A)Y(0) \\ &= E[Y(0)] + (Y(1) - Y(0))A + (Y(0) - E[Y(0)]) \\ &= \gamma_0 + \gamma_1 A + U \ , \end{aligned}$$

where

$$\begin{aligned} \gamma_0 &= E[Y(0)] \\ \gamma_1 &= Y(1) - Y(0) \\ U &= Y(0) - E[Y(0)] \ . \end{aligned}$$

If we let $X = (1, A)'$ and $\gamma = (\gamma_0, \gamma_1)'$, we obtain $Y = X'\gamma + U$ and this settings appears to fit the regression framework we have been discussing. However, there are some important differences. The most important one is that the potential outcomes are random variables, so $\gamma_1 = Y(1) - Y(0)$, the treatment effect, is also **random** - as opposed to being a constant unknown parameter. This captures the so-called *heterogenous treatment response*, which is the fact that the treatment effect may be different for different individuals. In order for this model to be exactly the linear regression model we discussed earlier we would need to make the additional assumption that the treatment effect is constant across individuals, that is,

$$Y_i(1) - Y_i(0) = c \quad \text{for all } i \leq n \ .$$

Under this, arguably strong, additional assumption, we can then write the linear causal model as $Y = X'\beta + U$ with β being a constant and $U \perp\!\!\!\perp X$ - since under random assignment the unobserved variable $U = Y(0) - E[Y(0)]$ is independent of the treatment assignment A. Notice that, in order to have a linear causal model a randomized controlled experiment is not enough; we also need a constant treatment effect.

Understanding the proper interpretation of the regression coefficients in a linear regression model is important due to its prevalence in applied work. Based on the previous derivation, one may conclude that the slope coefficient in a regression of Y on the treatment indicator variable A and a constant term would identify the *homogenous* treatment effect; that is, the causal effect of how the treatment affects the outcome for any possible individual. Under the homogeneity assumption such an interpretation would be correct, but unfortunately it is quite difficult to justify such an assumption in applications. Another approach is to allow for the treatment effects to be heterogeneous and try to characterize the so-called least squares *estimand* as a function of the treatment effect $Y(1) - Y(0)$. Under random assignment, we will show in the

next section that a regression of Y on A identifies the average treatment effect (ATE), essentially by the characterization we derived in (1.4). The ATE is a causal parameter because it is an *average* of causal effects.

2.3 Linear Regression when with Exogenous Regressors

2.3.1 Solving for β

Let (Y, X, U) be a random vector where Y and U take values in $\mathbf{R}$ and X takes values in $\mathbf{R}^{k+1}$. Assume further that the first component of X is a constant equal to one, i.e., $X = (X_0, X_1, \ldots, X_k)'$ with $X_0 = 1$. Let $\beta = (\beta_0, \beta_1, \ldots, \beta_k)' \in \mathbf{R}^{k+1}$ be such that

$$Y = X'\beta + U \ .$$

Suppose $E[XU] = 0$, $E[XX'] < \infty$, and that there is *no perfect collinearity in* X. The justification of the first assumption varies depending on which of the three preceding interpretations we invoke. The second assumption ensures that $E[XX']$ exists. The third assumption is equivalent to the assumption that the matrix $E[XX']$ is in fact invertible. Since $E[XX']$ is positive semi-definite, invertibility of $E[XX']$ is equivalent to $E[XX']$ being positive definite. We say that there is *perfect collinearity* or *multicollinearity* in X if there exists nonzero $c \in \mathbf{R}^{k+1}$ such that $P\{c'X = 0\} = 1$, i.e., if we can express one component of X as a linear combination of the others.

Lemma 2.1 *Let X be a random vector such that $E[XX'] < \infty$. Then $E[XX']$ is invertible if and only if there is no perfect collinearity in X.*

PROOF: We first argue that if $E[XX']$ is invertible, then there is no perfect collinearity in X. To see this, suppose there is perfect collinearity in X, i.e., that there exists a nonzero $c \in \mathbf{R}^{k+1}$ such that $P\{c'X = 0\} = 1$. Note that $E[XX']c = E[X(X'c)] = 0$. Hence, the columns of $E[XX']$ are linearly dependent, i.e., $E[XX']$ is not invertible.

We now argue that if there is no perfect collinearity in X, then $E[XX']$ is invertible. To see this, suppose $E[XX']$ is not invertible. Then, the columns of $E[XX']$ must be linearly dependent, i.e., there exists nonzero $c \in \mathbf{R}^{k+1}$ such that $E[XX']c = 0$. This implies further that $c'E[XX']c = E[(c'X)^2] = 0$, which in turn implies that $P\{c'X = 0\} = 1.$, i.e., that there is perfect collinearity in X. ■

The first assumption above together with the fact that $U = Y - X'\beta$ implies that $E[X(Y - X'\beta)] = 0$, i.e., $E[XY] = E[XX']\beta$. Since $E[XX']$ is invertible, we have that there is a unique solution to this system of equations, namely,

$$\beta = E[XX']^{-1}E[XY] \ . \tag{2.1}$$

If $E[XX']$ is not invertible, i.e., there is perfect collinearity in X, then there will be more than one solution to this system of equations. Importantly, any two solutions β and $\tilde{\beta}$ will necessarily satisfy $P\{X'\beta = X'\tilde{\beta}\} = 1$. Depending on the interpretation, this may be an important distinction or not. For instance, in the second interpretation, each such solution corresponds to the same "best" linear predictor of Y given X, whereas in the third interpretation different values of β could have wildly different implications for how X affects Y holding U constant.

2.3.2 Interpretation as the ATE

Now that we have a representation of β as a function of the distribution of (Y, X), we could go back to the case where $X = (1, A)'$ and A is a binary treatment satisfying $A \perp\!\!\!\perp (Y(1), Y(0))$. In this case, we may wonder what would the right interpretation for β be if we believe that the treatment effect is heterogeneous. In other words, we would like to characterize the least squares estimand as a function of the treatment effect $Y(1) - Y(0)$. To this end, let $\pi := E[A]$ and note that

$$\begin{bmatrix}\beta_0\\ \beta_1\end{bmatrix} = E\begin{bmatrix}1 & A\\ A & A\end{bmatrix}^{-1} E\begin{bmatrix}Y\\ AY\end{bmatrix} = \frac{1}{\pi(1-\pi)}\begin{bmatrix}\pi & -\pi\\ -\pi & 1\end{bmatrix}\begin{bmatrix}E[Y]\\ E[AY]\end{bmatrix} .$$

The results follows from simple matrix algebra. Next class we will learn tricks that will allow us to derive the same expression without inverting matrices at all, but for the moment we rely on the formula to invert a 2x2 matrix as a way to find the expression for β_1. In particular, we have that

$$\begin{aligned}\beta_1 &= \frac{E[AY] - \pi E[Y]}{\pi(1-\pi)}\\ &= \frac{E[Y \mid A=1] - E[Y]}{(1-\pi)}\\ &= \frac{(1-\pi)E[Y \mid A=1] - (1-\pi)E[Y \mid A=0]}{(1-\pi)}\\ &= E[Y(1)] - E[Y(0)] ,\end{aligned} \tag{2.2}$$

where the second equality follows from $E[YA] = \pi E[Y \mid A = 1]$ and the LIE, the third equality follows from $E[Y] = \pi E[Y \mid A = 1] + (1-\pi)E[Y \mid A = 0]$, and the last equality follows from random assignment. We conclude that the slope coefficient in a least square regression of Y on $(1, A)$ identifies $E[Y|A = 1] - E[Y|A = 0]$, which in tern equals the average treatment effect under random assignment. In other words, we can interpret the coefficient β_1 causally in the context of a linear regression when the treatment has been randomly generated by a randomized control experiment. This result breaks down very quickly (as soon as A takes more than two values in general), but it is important enough that it is worth highlighting.

A few considerations are important about this special case. The linear model in this case is not viewed as a linear causal model (as, for example, in interpretation 3 above). Instead, the linear model is viewed as a convenient way to summarize the joint distribution of (Y, X)(as in interpretation 2 above), and then it happens to be that this summary statistic (the slope coefficient) is equal to the ATE when the treatment variable A is independently assigned to units. Broadly speaking, the linear model with homogenous partial effects is rarely seen as the "true" causal model, but most often the focus is on trying to understand when this estimand, i.e., β as in (2.1), can be written as a function (often a weighted average) of well-defined causal effects. To gain some appreciation about this last statement, suppose that the true causal model is $Y = X'\gamma + U$ with $U \perp\!\!\!\perp X$ and γ being random, so the partial effect of X on Y is actually given by γ but the model allows for each individual to respond differently to changes in X. That is, if we index units by i, the coefficient γ would become γ_i. In this case, it is immediate to show that β in (2.1) equals

$$\beta = E[\omega'\gamma] \quad \text{with} \quad \omega' = E[XX']^{-1}XX' \ .$$

It follows that each element of β is a weighted average of the vector γ. However, even though these weights satisfy $E[\omega] = \mathbb{I}$, they may be negative with positive probability and therefore prevent us from interpreting β as a convex weighted average of the causal effects γ in general. For example, it becomes possible that $\gamma > 0$ with probability one and yet $\beta < 0$. We will go back to this point later in class as we introduce additional concepts.

2.4 Estimating β

Let (Y, X, U) be a random vector where Y and U take values in $\mathbf{R}$ and X takes values in $\mathbf{R}^{k+1}$. Assume further that the first component of X is a constant equal to one. Let $\beta \in \mathbf{R}^{k+1}$ be such that

$$Y = X'\beta + U \ .$$

Suppose that $E[XU] = 0$, $E[XX'] < \infty$ and that there is no perfect collinearity in X. Above we described three different interpretations and justifications of such a model. We now discuss estimation of β.

2.4.1 Ordinary Least Squares

Let (Y, X, U) be distributed as described above and denote by P the marginal distribution of (Y, X). Let $(Y_1, X_1), \ldots, (Y_n, X_n)$ be an i.i.d. sequence of random vectors with distribution P. By analogy with the expression we derived

for β under these assumptions, the natural estimator of β is simply

$$\hat{\beta}_n = \left(\frac{1}{n} \sum_{1 \leq i \leq n} X_i X_i' \right)^{-1} \left(\frac{1}{n} \sum_{1 \leq i \leq n} X_i Y_i \right) .$$

This estimator is called the *ordinary least squares* (OLS) estimator of β because it can also be derived as the solution to the following minimization problem:

$$\min_{b \in \mathbf{R}^{k+1}} \frac{1}{n} \sum_{1 \leq i \leq n} (Y_i - X_i' b)^2 .$$

To see this, note that $\frac{1}{n} \sum_{1 \leq i \leq n} (Y_i - X_i' b)^2$ is convex (as a function of b) and

$$D_b \frac{1}{n} \sum_{1 \leq i \leq n} (Y_i - X_i' b)^2 = \frac{1}{n} \sum_{1 \leq i \leq n} -2 X_i (Y_i - X_i' b) .$$

Hence $\hat{\beta}_n$ must satisfy

$$\frac{1}{n} \sum_{1 \leq i \leq n} X_i (Y_i - X_i' \hat{\beta}_n) = 0 ,$$

i.e.,

$$\frac{1}{n} \sum_{1 \leq i \leq n} X_i Y_i = \frac{1}{n} \sum_{1 \leq i \leq n} X_i X_i' \hat{\beta}_n .$$

The matrix

$$\frac{1}{n} \sum_{1 \leq i \leq n} X_i X_i'$$

may not be invertible, but, since $E[XX']$ is invertible, it will be invertible with probability approaching one.

The ith *fitted value* is denoted by $\hat{Y}_i = X_i' \hat{\beta}_n$. The ith *residual* is denoted by $\hat{U}_i = Y_i - \hat{Y}_i$. By definition, we therefore have that

$$\frac{1}{n} \sum_{1 \leq i \leq n} X_i \hat{U}_i = 0 .$$

2.4.2 Projection Interpretation

Define

$$\begin{aligned}
\mathbb{Y} &= (Y_1, \ldots, Y_n)' \\
\mathbb{X} &= (X_1, \ldots, X_n)' \\
\hat{\mathbb{Y}} &= (\hat{Y}_1, \ldots, \hat{Y}_n)' \\
&= \mathbb{X}\hat{\beta}_n \\
\mathbb{U} &= (U_1, \ldots, U_n)' \\
\hat{\mathbb{U}} &= (\hat{U}_1, \ldots, \hat{U}_n)' \\
&= \mathbb{Y} - \hat{\mathbb{Y}} \\
&= \mathbb{Y} - \mathbb{X}\hat{\beta}_n \ .
\end{aligned}$$

In this notation,

$$\hat{\beta}_n = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbb{Y}$$

and may be equivalently described as the solution to

$$\min_{b \in \mathbf{R}^{k+1}} |\mathbb{Y} - \mathbb{X}b|^2 \ .$$

Hence, $\mathbb{X}\hat{\beta}_n$ is the vector in the column space of $\mathbb{X}$ that is closest (in terms of Euclidean distance) to $\mathbb{Y}$. From the above, we see that $\mathbb{X}'\hat{\mathbb{U}} = 0$, thus $\hat{\mathbb{U}}$ is orthogonal to all of the columns of $\mathbb{X}$ (and thus orthogonal to all of the vectors in the column space of $\mathbb{X}$). In this sense,

$$\mathbb{X}\hat{\beta}_n = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbb{Y}$$

is the *orthogonal projection* of $\mathbb{Y}$ onto the ($(k+1)$-dimensional) column space of $\mathbb{X}$. The matrix

$$\mathbb{P} = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'$$

is known as a *projection* matrix. It projects a vector in $\mathbf{R}^n$ (such as $\mathbb{Y}$) onto the column space of $\mathbb{X}$. Note that $\mathbb{P}^2 = \mathbb{P}$, which reflects the fact that projecting something that already lies in the column space of $\mathbb{X}$ onto the column space of $\mathbb{X}$ does nothing. The matrix $\mathbb{P}$ is also symmetric. The matrix

$$\mathbb{M} = \mathbb{I} - \mathbb{P}$$

is also a projection matrix. It projects a vector onto the ($(n-k-1)$-dimensional) vector space orthogonal to the column space of $\mathbb{X}$. Hence, $\mathbb{M}\mathbb{X} = 0$. Note that $\mathbb{M}\mathbb{Y} = \hat{\mathbb{U}}$. For this reason, $\mathbb{M}$ is sometimes called the "residual maker" matrix.

2.5 Concluding Remarks

These notes are based on past notes for Econ 480-3 that I have written for previous editions of this class, and are heavily influenced by Azeem Shaikh, the notes he kindly shared with me, and our conversations about teaching over the years. Other sources that contain useful related concepts include the books by Bruce Hansen, Hansen [2022], and the one by Angrist and Pischke, Angrist and Pischke [2008].

2.6 Problems

Problem 2.1 *Let (Y, X) be a random vector taking values in $\mathbf{R}^2$ with finite first and second moments.*

a) Show that, without loss of generality, we can write

$$Y = h(X) + U \ ,$$

where U is a scalar random variable satisfying $E[U|X] = 0$ and $h(X)$ is a function of X.

b) Given an interpretation to $h(X)$.

Problem 2.2 *Prove the assertion in the last paragraph of Section 2.3.1: If $E[XX']$ is not invertible, i.e., there is perfect collinearity in X, then there will be more than one solution to this system of equations. Importantly, any two solutions β and $\tilde{\beta}$ will necessarily satisfy $P\{X'\beta = X'\tilde{\beta}\} = 1$.*

Problem 2.3 *Prove the result in Section 2.3.2 differently by showing $\beta_1 = \frac{\text{Cov}[Y,A]}{\text{Var}[A]}$ and $\text{Cov}[Y, A] = \mathbb{E}[Y(1) - Y(0)]\,\text{Var}[A]$.*

Problem 2.4 *Show that for any vector $\mathbb{W} \in \mathbf{R}^n$, $\mathbb{M}\mathbb{W}$ is orthogonal to any vector in the column space of $\mathbb{X}$. That is, for any $b \in \mathbf{R}^{k+1}$,*

$$(\mathbb{X}b)'(\mathbb{M}\mathbb{W}) = 0.$$

Bibliography

J. D. Angrist and J.-S. Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2008.

B. Hansen. *Econometrics*. Princeton University Press, 2022.

3

More on Linear Regression

Write intro to this lecture hre.

3.1 Solving for Sub-vectors of β

Partition X into X_1 and X_2, where X_1 takes values in $\mathbf{R}^{k_1}$ and X_2 takes values in $\mathbf{R}^{k_2}$. Partition β into β_1 and β_2 analogously. In this notation,

$$Y = X_1'\beta_1 + X_2'\beta_2 + U \ .$$

Our preceding results imply that

$$\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} E[X_1X_1'] & E[X_1X_2'] \\ E[X_2X_1'] & E[X_2X_2'] \end{pmatrix}^{-1} \begin{pmatrix} E[X_1Y] \\ E[X_2Y] \end{pmatrix} \ .$$

Using the so called partitioned matrix inverse formula, it would be possible to derive formulae for β_1 and β_2, but such an exercise is not particularly illuminating. We therefore take a different approach to arrive at the same formulae. In doing so, we will make use of the following notation: for a random variable A and a random vector B, denote by $\text{BLP}(A|B)$ the best linear predictor of A given B, i.e. $B'E[BB']^{-1}E[BA]$. Recall that the best linear predictor is the second interpretation of the linear regression and here we use the formula for the regression coefficients derived previously. If A is a random vector, then define $\text{BLP}(A|B)$ component-wise.

Define $\tilde{Y} = Y - \text{BLP}(Y|X_2)$ and $\tilde{X}_1 = X_1 - \text{BLP}(X_1|X_2)$. Consider the linear regression $\tilde{Y} = \tilde{X}_1'\tilde{\beta}_1 + \tilde{U}$, where $E[\tilde{X}_1\tilde{U}] = 0$ (again, using the second interpretation of the linear regression model). It follows that $\tilde{\beta}_1 = \beta_1$. To see this, note that $E[\tilde{X}_1\tilde{X}_1']$ is invertible (because each component of $\tilde{X}_1$ is a linear combination the components of X), so

$$\begin{aligned} \tilde{\beta}_1 &= E[\tilde{X}_1\tilde{X}_1']^{-1}E[\tilde{X}_1\tilde{Y}] \\ &= E[\tilde{X}_1\tilde{X}_1']^{-1}(E[\tilde{X}_1Y] - E[\tilde{X}_1\text{BLP}(Y|X_2)] \\ &= E[\tilde{X}_1\tilde{X}_1']^{-1}E[\tilde{X}_1Y] \ , \end{aligned}$$

where the first equality follows from the formula for $\tilde{\beta}_1$, the second equality

follows from the expression for $\tilde{Y}$, and the third equality follows from the fact that $E[\tilde{X}_1 X_2'] = 0$ (because $\tilde{X}_1$ is the error term from a regression of X_1 on X_2). Note that this first part of the derivation shows that $\tilde{\beta}_1$ is also the population coefficient of a linear regression of Y on $\tilde{X}_1$. If we now replace Y by its expression and do some additional steps, we get

$$\begin{aligned}\tilde{\beta}_1 &= E[\tilde{X}_1\tilde{X}_1']^{-1}(E[\tilde{X}_1 X_1'\beta_1] + E[\tilde{X}_1 X_2'\beta_2] + E[\tilde{X}_1 U]) \\ &= E[\tilde{X}_1\tilde{X}_1']^{-1}(E[\tilde{X}_1 X_1'\beta_1]) \\ &= E[\tilde{X}_1\tilde{X}_1']^{-1}(E[\tilde{X}_1\tilde{X}_1'\beta_1] + E[\tilde{X}_1 \mathrm{BLP}(X_1|X_2)'\beta_1]) \\ &= \beta_1 \ ,\end{aligned}$$

where the first equality follows from the expression for Y, the second equality follows from the fact that $E[\tilde{X}_1 X_2'] = 0$ and $E[\tilde{X}_1 U] = 0$ (because $E[XU] = 0$), the third equality follows from the expression for $\tilde{X}_1$, and the final equality follows from the fact that $E[\tilde{X}_1 X_2'] = 0$.

In other words, β_1 in the linear regression of Y on X_1 and X_2 is equal to the coefficient in a linear regression of the error term from a linear regression of Y on X_2 on the error terms from a linear regression of the components of X_1 on X_2. This gives meaning to the common description of β_1 as the "effect" of X_1 on Y after "controlling for X_2."

Notice that if we take X_2 to be just a constant, then $\tilde{Y} = Y - E[Y]$ and $\tilde{X}_1 = X_1 - E[X_1]$. Hence,

$$\begin{aligned}\beta_1 &= E[(X_1 - E[X_1])(X_1 - E[X_1])']^{-1}E[(X_1 - E[X_1])(Y - E[Y])] \\ &= \mathrm{Var}[X_1]^{-1}\mathrm{Cov}[X_1, Y] \ .\end{aligned}$$

Finally, also note that if we use our formula to interpret the coefficient β_j associated with the jth covariate for $1 \leq j \leq k$, we obtain

$$\beta_j = \frac{\mathrm{Cov}[\tilde{X}_j, Y]}{\mathrm{Var}[\tilde{X}_j]} \ , \tag{3.1}$$

which shows that each coefficient in a multivariate regression is the bivariate slope coefficient for the corresponding regressor, after "partialling out" all the other variables in the model.

3.2 Estimating Sub-Vectors of β

Partition X into X_1 and X_2, where X_1 takes values in $\mathbf{R}^{k_1}$ and X_2 takes values in $\mathbf{R}^{k_2}$. Partition β into β_1 and β_2 analogously. In this notation,

$$Y = X_1'\beta_1 + X_2'\beta_2 + U \ .$$

Using the preceding results, we can derive estimation counterparts to the results above about solving for sub-vectors of β. Again, this may be done using the partitioned matrix inverse formula, but we will use a different approach. Let $\mathbb{X}_1 = (X_{1,1}, \ldots, X_{1,n})'$ and $\mathbb{X}_2 = (X_{2,1}, \ldots, X_{2,n})'$. Denote by $\mathbb{P}_1$ the projection matrix onto the column space of $\mathbb{X}_1$ and $\mathbb{P}_2$ the projection matrix onto the column space of $\mathbb{X}_2$. Define $\mathbb{M}_1 = \mathbb{I} - \mathbb{P}_1$ and $\mathbb{M}_2 = \mathbb{I} - \mathbb{P}_2$.

First note that

$$\mathbb{Y} = \mathbb{X}_1\hat{\beta}_{1,n} + \mathbb{X}_2\hat{\beta}_{2,n} + \hat{\mathbb{U}} \; .$$

This implies that

$$\mathbb{M}_2\mathbb{Y} = \mathbb{M}_2\mathbb{X}_1\hat{\beta}_{1,n} + \hat{\mathbb{U}}$$

because $\mathbb{M}_2\mathbb{X}_2 = 0$ and $\mathbb{M}_2\hat{\mathbb{U}} = \hat{\mathbb{U}}$, as $\hat{\mathbb{U}}$ is orthogonal to the column space of $\mathbb{X}$ (and hence the column space of $\mathbb{X}_2$ as well). This implies further that

$$(\mathbb{M}_2\mathbb{X}_1)'\mathbb{M}_2\mathbb{Y} = (\mathbb{M}_2\mathbb{X}_1)'\mathbb{M}_2\mathbb{X}_1\hat{\beta}_{1,n}$$

because $(\mathbb{M}_2\mathbb{X}_1)'\hat{\mathbb{U}} = \mathbb{X}_1'\mathbb{M}_2\hat{\mathbb{U}} = \mathbb{X}_1'\hat{\mathbb{U}} = 0$, as $\mathbb{X}'\hat{\mathbb{U}} = 0$. Note that the matrix $(\mathbb{M}_2\mathbb{X}_1)'(\mathbb{M}_2\mathbb{X}_1)$ is invertible provided that $\mathbb{X}'\mathbb{X}$ is invertible. Hence,

$$\hat{\beta}_{1,n} = ((\mathbb{M}_2\mathbb{X}_1)'(\mathbb{M}_2\mathbb{X}_1))^{-1}(\mathbb{M}_2\mathbb{X}_1)'\mathbb{M}_2\mathbb{Y} \; .$$

In other words, $\hat{\beta}_{1,n}$ can be obtained by estimating via OLS the coefficients from a linear regression of $\mathbb{M}_2\mathbb{Y}$ on $\mathbb{M}_2\mathbb{X}_1$. Upon recognizing that $\mathbb{M}_2\mathbb{Y}$ are the residuals from a regression of $\mathbb{Y}$ on $\mathbb{X}_2$ and that the columns of $\mathbb{M}_2\mathbb{X}_1$ are the residuals from regressions of the columns of $\mathbb{X}_1$ on $\mathbb{X}_2$, we see that this formula exactly parallels the formula we derived earlier for a sub-vector of β. This result is sometimes referred to as the *Frisch-Waugh-Lovell* (FWL) decomposition.

3.3 Application to Saturated in Discrete Covariates

Let's now consider a special case of the subvector problem that often arises in certain empirical settings. Suppose we can partition X into (A, W), where A is a binary treatment of interest and W are discrete covariates taking values in the set $\mathcal{W}$.

Example 3.1 (Angrist (1998)) Angrist (98) Angrist [1998] studies the causal effects of voluntary military services in the US on the later earning of soldiers. In this application, Y is a labor market outcome like employment or earnings, the treatment A denotes veteran status (i.e., participation in the military), and W includes socioeconomic variables like race, year of birth, schooling, application year, and the Armed Forces Qualification Test (AFQT) scores. Angrist (1998) shows that the coefficient on A in a regression that is

"saturated" in W, admits the interpretation of a weighted average of conditional average treatment effects (under the assumption that A is independent of potential earnings conditional on all of these covariates; an assumption we formally introduce in the next lecture). This result follows from the FWL decomposition as we discuss next. ■

Let $I_w := I\{W = w\}$ denote the indicator for the covariate W being equal to w. Consider the regression

$$Y = \beta_A A + \sum_{w \in \mathcal{W}} \gamma_w I_w + U \ , \tag{3.2}$$

where we have now partitioned the coefficient β into β_A and $\{\gamma_w : w \in \mathcal{W}\}$. The parameter of interest in this regression is β_A, and while we interpret this parameter as a projection coefficient (interpretation 2), the hope is that we can show that it captures some form of treatment effect of A on Y (a feature that will depend on the assumptions we impose in the model). The coefficients γ_w are viewed as nuisance regression parameters associated with each value of the covariates W. Note that this regression model allows for a separate parameter, γ_w, for every value taken on by the covariates. This model can therefore be said to be saturated-in-W, since it includes a parameter for every value of W. This is not the same as including W linearly, which would include the term $W'\gamma$. It is also not a "fully"-saturated model because γ_w is not indexed by a (i.e., there is no interaction between A and I_w). A fully saturated model would look as follows,

$$Y = \sum_{w \in \mathcal{W}} \gamma_w^* I_w + \sum_{w \in \mathcal{W}} \delta_w^* A I_w + U^* \ . \tag{3.3}$$

We can use the results on representation of subvectors we just derived to obtain an interesting characterization of the parameter β_A. To see this, start with

$$\beta_A = E[\tilde{A}^2]^{-1} E[\tilde{A} Y] \quad \text{where} \quad \tilde{A} = A - \text{BLP}(A | \{I_w\}_{w \in \mathcal{W}}) \ . \tag{3.4}$$

In addition, since W is discrete and $\{I_w\}_{w \in \mathcal{W}}$ is a set of indicators of all possible values that W can take, then it follows that if we let $\pi(W) := P\{A = 1 \mid W\}$,

$$\text{BLP}(A | I_w) = E[A|W] = \pi(W) \Rightarrow \tilde{A} = A - \pi(W) \ , \tag{3.5}$$

see Problem 3.1. By the LIE, we also have that $E[\tilde{A} Y] = E[\tilde{A} E[Y \mid A, W]]$, and so we can write

$$E[Y \mid A, W] = E[Y \mid A = 0, W] + A \Delta(W) \ , \tag{3.6}$$

where

$$\Delta(W) := E[Y \mid A = 1, W] - E[Y \mid A = 0, W] \ .$$

Race	Average earnings in 1988-1991	Differences in means by veteran status	Matching estimates	Regression estimates	Regression minus matching
	(1)	(2)	(3)	(4)	(5)
Whites	14537	1233.4	-197.2	-88.8	108.4
		(60.3)	(70.5)	(62.5)	(28.5)
Non-whites	11664	2449.1	839.7	1074.4	234.7
		(47.4)	(62.7)	(50.7)	(32.5)

TABLE 3.1: Uncontrolled, matching, and regression estimates of the effects of voluntary military service on earnings. Source: MHE p.73.

Finally, consider the following derivation

$$\begin{aligned} E[\tilde{A}Y] &= E[\tilde{A}E[Y \mid A, W]] \\ &= E[\tilde{A}E[Y \mid A = 0, W]] + E[\tilde{A}A\Delta(W)] \\ &= E[\tilde{A}A\Delta(W)] \\ &= E[\tilde{A}^2\Delta(W)] + E[\tilde{A}\pi(W)\Delta(W)] \\ &= E[\tilde{A}^2\Delta(W)] \end{aligned} \tag{3.7}$$

where the third and fifth equality follows from $E[\tilde{A} \mid W] = 0$ and the LIE, and the fourth equality follows by adding and subtracting $\pi(W)$. Using this last expression back into (3.4), we obtain

$$\beta_A = E\left[\frac{\tilde{A}^2}{E[\tilde{A}^2]}\Delta(W)\right] , \tag{3.8}$$

which shows that β_A is a weighted average of $\Delta(W)$. Under the type of selection on observables assumptions we will discuss next class, it turns out that $\Delta(W)$ can be interpreted as a conditional average treatment effect (CATE) and so β_A is a weighted averages of CATEs. The weights in this characterization are random (by virtue of $\tilde{A}$ being random), but there is an alternative characterization of β_A with non-random weights; see Problem A.3. Unless we impose further, and usually strong, assumptions, the estimand β_A is not equal to the ATT or ATE.

Column (4) in Table 3.1 present estimated values of β_A for each race, as those reports in Angrist [1998]. We will discuss how to interpret the numbers in the different columns next class. For now, recall that W included variables such as race, year of birth, schooling, application year, and the AFQT score. The indicator I_w then indicates each possible combination of values that the vector W can take. To get a sense of the dimensionality in Table 3.1, conditional on

race, there are approximately 450 possible values of W and roughly $150K$ observations. Note how race is treated as a special covariate, since the value of β_A is allowed to vary by race but not by the rest of the characteristics of the veterans.

3.4 Covariance Adjustment under Random Assignment

Suppose now that $A \in \{0,1\}$ is randomly assigned with $P\{A=1\} = \pi \in (0,1)$, and let W denote baseline covariates. Let

$$\theta := E[Y(1) - Y(0)]$$

denote the average treatment effect (ATE). In a randomized experiment, the target θ is identified by random assignment alone. However, in many settings, researchers have access to additional information from covariates (such as age, gender, etc.) and it is natural to ask whether we can use this information to improve the precision of our ATE estimates. Is it possible to do better than the difference-in-means estimator? If so, how?

It turns out that while the difference-in-means estimator is simple and effective, incorporating covariates can potentially yield more accurate results, especially when these covariates are correlated with the outcome of interest. It is important to keep in mind that "covariate adjustment" is not about "fixing bias" from confounding in the context of an RCT. Instead, adjustment is about precision. Intuitively, if part of the variation in Y is predictable from baseline variables W, then removing that predictable component before taking treated-versus-control differences can reduce noise. In this chapter we will try to formalize this intuition.

To make the arguments clean, assume that $E[W] = 0$ without loss of generality, as we could always recenter the covariates otherwise. Then, for any conformable vectors γ_1 and γ_0, define centered arm-specific residualized outcomes

$$Y^{\gamma_1} := Y - W'\gamma_1 \quad \text{for } A = 1 \ ,$$
$$Y^{\gamma_0} := Y - W'\gamma_0 \quad \text{for } A = 0 \ ,$$

and consider

$$\Delta(\gamma_1, \gamma_0) := E[Y^{\gamma_1} \mid A = 1] - E[Y^{\gamma_0} \mid A = 0] \ . \tag{3.9}$$

Proposition 3.1 *Assume $A \perp\!\!\!\perp (Y(1), Y(0), W)$. Then, for every pair (γ_1, γ_0),*

$$\Delta(\gamma_1, \gamma_0) = \theta \ .$$

PROOF. Using consistency, $Y = AY(1) + (1 - A)Y(0)$. Random assignment implies

$$E[Y \mid A = 1] = E[Y(1)], \qquad E[Y \mid A = 0] = E[Y(0)],$$

and $E[W] = E[W \mid A = 1] = E[W \mid A = 0] = 0$. Therefore

$$\Delta(\gamma_1, \gamma_0) = E[Y(1)] - E[Y(0)] = \theta \ .$$

■

The proposition formalizes the idea that covariate adjustment is not about changing the target parameter. No matter how we choose (γ_1, γ_0), the population quantity $\Delta(\gamma_1, \gamma_0)$ remains equal to the ATE. Thus, the role of covariate adjustment in a randomized experiment is not to remove bias, but *potentially* to improve precision.

To see why the choice of (γ_1, γ_0) matters, consider the natural sample analogue of $\Delta(\gamma_1, \gamma_0)$:

$$\begin{aligned}\hat{\Delta}_n(\gamma_1, \gamma_0) &:= \frac{1}{n_1} \sum_{i \in \mathcal{I}_1} (Y_i - W_i'\gamma_1) - \frac{1}{n_0} \sum_{i \in \mathcal{I}_0} (Y_i - W_i'\gamma_0) \\ &= \frac{1}{n_1} \sum_{i=1}^{n} (Y_i - W_i'\gamma_1) A_i - \frac{1}{n_0} \sum_{i=1}^{n} (Y_i - W_i'\gamma_0)(1 - A_i) \ . \end{aligned} \tag{3.10}$$

Since treatment is randomly assigned, $\hat{\Delta}_n(\gamma_1, \gamma_0)$ is a consistent estimator of θ for every choice of (γ_1, γ_0). Therefore, the relevant question is whether different choices of (γ_1, γ_0) lead to estimators with different precision.

Because (3.10) is a difference in sample means, its asymptotic variance is determined by the conditional variances of the residualized outcomes in the treatment and control groups. In particular, under standard regularity conditions,

$$\sqrt{n}\big(\hat{\Delta}_n(\gamma_1, \gamma_0) - \theta\big) \xrightarrow{d} N\big(0, V(\gamma_1, \gamma_0)\big) \ , \tag{3.11}$$

where

$$V(\gamma_1, \gamma_0) := \frac{\mathrm{Var}(Y - W'\gamma_1 \mid A = 1)}{\pi} + \frac{\mathrm{Var}(Y - W'\gamma_0 \mid A = 0)}{1 - \pi} \ .$$

See Problem 3.6 for a derivation. This expression is intuitive. The treated-group average is based on approximately $n\pi$ observations, so its contribution to the asymptotic variance is the treated-arm variance divided by π. Likewise, the control-group average is based on approximately $n(1-\pi)$ observations, so its contribution is the control-arm variance divided by $1-\pi$. The total asymptotic variance is the sum of these two components. Thus, choosing (γ_1, γ_0) amounts to choosing how to residualize outcomes within each treatment arm so as to make these conditional variances as small as possible.

Proposition 3.2 (Optimal arm-specific linear adjustment) *Let*

$$\Lambda_1 := \operatorname{Cov}(Y, W \mid A = 1) \ , \qquad \Lambda_0 := \operatorname{Cov}(Y, W \mid A = 0) \ ,$$

$$Q_1 := \operatorname{Var}(W \mid A = 1) \ , \qquad Q_0 := \operatorname{Var}(W \mid A = 0) \ .$$

Then

$$V(\gamma_1, \gamma_0) = V(0,0) - \frac{2}{\pi}\gamma_1'\Lambda_1 + \frac{1}{\pi}\gamma_1' Q_1 \gamma_1 - \frac{2}{1-\pi}\gamma_0'\Lambda_0 + \frac{1}{1-\pi}\gamma_0' Q_0 \gamma_0 \ .$$

If Q_1 and Q_0 are invertible, the unique minimizers are

$$\gamma_1^\star = Q_1^{-1}\Lambda_1, \qquad \gamma_0^\star = Q_0^{-1}\Lambda_0 \ .$$

Proof. By definition,

$$V(\gamma_1, \gamma_0) = \frac{\operatorname{Var}(Y - W'\gamma_1 \mid A = 1)}{\pi} + \frac{\operatorname{Var}(Y - W'\gamma_0 \mid A = 0)}{1-\pi} \ .$$

For each $a \in \{0, 1\}$,

$$\operatorname{Var}(Y - W'\gamma_a \mid A = a) = \operatorname{Var}(Y \mid A = a) - 2\gamma_a' \operatorname{Cov}(Y, W \mid A = a) + \gamma_a' \operatorname{Var}(W \mid A = a)\gamma_a \ .$$

Applying this identity for $a = 1$ and $a = 0$ yields

$$\begin{aligned} V(\gamma_1, \gamma_0) = {} & \frac{\operatorname{Var}(Y \mid A = 1)}{\pi} - \frac{2}{\pi}\gamma_1'\Lambda_1 + \frac{1}{\pi}\gamma_1' Q_1 \gamma_1 \\ & + \frac{\operatorname{Var}(Y \mid A = 0)}{1-\pi} - \frac{2}{1-\pi}\gamma_0'\Lambda_0 + \frac{1}{1-\pi}\gamma_0' Q_0 \gamma_0 \ . \end{aligned}$$

Since

$$V(0,0) = \frac{\operatorname{Var}(Y \mid A = 1)}{\pi} + \frac{\operatorname{Var}(Y \mid A = 0)}{1-\pi} \ ,$$

the first claim follows.

To minimize $V(\gamma_1, \gamma_0)$, note that the objective is separable in γ_1 and γ_0, so each coefficient vector can be chosen independently. For the treated arm, we minimize

$$-\frac{2}{\pi}\gamma_1'\Lambda_1 + \frac{1}{\pi}\gamma_1' Q_1 \gamma_1 \ .$$

Because Q_1 is invertible, this is a strictly convex quadratic function of γ_1. Its first-order condition is

$$-\frac{2}{\pi}\Lambda_1 + \frac{2}{\pi}Q_1\gamma_1 = 0 \ ,$$

which implies

$$\gamma_1^\star = Q_1^{-1}\Lambda_1 \ .$$

The same argument for the control arm gives

$$\gamma_0^\star = Q_0^{-1}\Lambda_0 \ .$$

Uniqueness follows from strict convexity. ■

Remark 3.1 (Intuition for $\gamma_1^\star$ and $\gamma_0^\star$) Since $E[W] = 0$, the coefficients $\gamma_1^\star$ and $\gamma_0^\star$ are the population least-squares slopes from regressing Y on W separately within each treatment arm:

$$\gamma_1^\star = \arg\min_\gamma E[(Y - W'\gamma)^2 \mid A = 1], \qquad \gamma_0^\star = \arg\min_\gamma E[(Y - W'\gamma)^2 \mid A = 0].$$

Thus, each coefficient vector captures the best linear predictor of Y from W within its arm. If W is strongly predictive of outcomes among treated units, then $\gamma_1^\star$ removes substantial noise in the treated sample; similarly, $\gamma_0^\star$ removes substantial noise in the control sample. ■

Corollary 3.1 (No-worse efficiency) *Under the assumptions of the previous proposition,*

$$V(\gamma_1, \gamma_0) - V(\gamma_1^\star, \gamma_0^\star) = \frac{1}{\pi}(\gamma_1 - \gamma_1^\star)' Q_1 (\gamma_1 - \gamma_1^\star) + \frac{1}{1-\pi}(\gamma_0 - \gamma_0^\star)' Q_0 (\gamma_0 - \gamma_0^\star) \geq 0 .$$

Hence

$$V(\gamma_1^\star, \gamma_0^\star) \leq V(0, 0).$$

Equality holds if and only if $\gamma_1^\star = 0$ and $\gamma_0^\star = 0$.

Remark 3.2 The corollary compares the oracle choice $(\gamma_1^\star, \gamma_0^\star)$ to the unadjusted estimator. It does *not* imply that

$$V(\gamma_1, \gamma_0) \leq V(0, 0)$$

for every pair (γ_1, γ_0). If the adjustment coefficients are chosen poorly, covariate adjustment can increase asymptotic variance rather than decrease it. ■

The pair $(\gamma_1^\star, \gamma_0^\star)$ is therefore an ideal "oracle" choice: it delivers the lowest possible asymptotic variance within this class, but it depends on unknown population quantities. In practice we have to replace $(\gamma_1^\star, \gamma_0^\star)$ by data-driven estimates. The key point is that as long as our estimators $(\hat{\gamma}_{1,n}, \hat{\gamma}_{0,n})$ are consistent for $(\gamma_1^\star, \gamma_0^\star)$—for example, if we estimate each slope by OLS within treatment arm—then the resulting adjusted estimator of θ has the same first-order asymptotic variance as the oracle-adjusted estimator. In other words, we can approximate the oracle efficiency bound arbitrarily well using feasible, sample-based adjustment.

Proposition 3.3 *Let $(\hat{\gamma}_{1,n}, \hat{\gamma}_{0,n})$ be estimators such that*

$$\hat{\gamma}_{1,n} \xrightarrow{p} \gamma_1^\star , \qquad \hat{\gamma}_{0,n} \xrightarrow{p} \gamma_0^\star .$$

Then

$$\sqrt{n}\Big(\hat{\Delta}_n(\hat{\gamma}_{1,n}, \hat{\gamma}_{0,n}) - \theta\Big) \xrightarrow{d} N\big(0, V(\gamma_1^\star, \gamma_0^\star)\big) .$$

Thus, replacing the oracle coefficients by any consistent estimators does not affect the first-order asymptotic variance.

The discussion in Remark 3.1 suggests a natural way to estimate $(\gamma_1^\star, \gamma_0^\star)$: run separate OLS regressions of Y on W within each treatment arm. Formally,

$$\hat{\gamma}_{1,n} = \arg\min_\gamma \frac{1}{n_1} \sum_{i \in \mathcal{I}_1} (Y_i - W_i'\gamma)^2 \ , \qquad \hat{\gamma}_{0,n} = \arg\min_\gamma \frac{1}{n_0} \sum_{i \in \mathcal{I}_0} (Y_i - W_i'\gamma)^2 \ .$$

Note that in practice W is not mean-zero, so we either recenter the covariates or include an intercept in each regression. If we include an intercept, the **two separate** least squares regressions we run woudl be :

- **regression 1**: LS of Y on $(1, W)$ for units with $A_i = 1$
- **regression 2**: LS of Y on $(1, W)$ for units with $A_i = 0$

It is important to emphasize that these are two separate regressions, not a single pooled regression of Y on $(1, A, W)$ with a common slope on W. The reason is that the efficiency bound developed above allows the best linear adjustment to differ across treatment arms. In general, the slope that minimizes asymptotic variance when we impose a common coefficient on W is not the same as the pair $(\gamma_1^\star, \gamma_0^\star)$ obtained from arm-specific adjustment. Moreover, even within the common-slope class, the variance-minimizing coefficient is generally not the same as the slope from the pooled regression of Y on $(1, W)$. Problem 3.5 makes this point precise and shows that, in general, the common-slope oracle differs from the pooled least-squares slope.

At the same time, the optimal arm-specific adjustment has a convenient regression implementation. Consider the population linear projection of Y on $(1, A, W, AW)$:

$$Y = \beta_0 + A\beta_1 + W'\beta_2 + AW'\beta_3 + U \ , \tag{3.12}$$

where the coefficients $(\beta_0, \beta_1, \beta_2, \beta_3)$ are LS projection coefficients (as in interpretation 2). Under random assignment and the normalization $E[W] = 0$, one can show that the projection coefficients satisfy

$$\beta_2 = \gamma_0^\star, \qquad \beta_3 = \gamma_1^\star - \gamma_0^\star, \qquad \beta_1 = \theta \ .$$

Thus, the coefficient on A in the fully interacted regression coincides with the ATE, and the coefficients on W and AW recover the optimal arm-specific adjustment slopes. In this sense, the interacted regression provides a convenient OLS implementation of the oracle adjustment discussed above.

An intuitive way to see this is as follows. Let

$$\delta_1^\star := E[Y(1) - W'\gamma_1^\star] \ , \qquad \delta_0^\star := E[Y(0) - W'\gamma_0^\star] \ ,$$

and define the residual potential outcomes

$$\epsilon_1^\star := Y(1) - W'\gamma_1^\star - \delta_1^\star \ , \qquad \epsilon_0^\star := Y(0) - W'\gamma_0^\star - \delta_0^\star \ .$$

Then

$$Y(1) = \delta_1^\star + W'\gamma_1^\star + \epsilon_1^\star \ , \qquad Y(0) = \delta_0^\star + W'\gamma_0^\star + \epsilon_0^\star \ .$$

Using consistency,

$$Y = AY(1) + (1 - A)Y(0) \ ,$$

so substituting the previous expressions yields

$$\begin{aligned} Y &= A\Big(\delta_1^\star + W'\gamma_1^\star + \epsilon_1^\star\Big) + (1 - A)\Big(\delta_0^\star + W'\gamma_0^\star + \epsilon_0^\star\Big) \\ &= \delta_0^\star + A(\delta_1^\star - \delta_0^\star) + W'\gamma_0^\star + AW'(\gamma_1^\star - \gamma_0^\star) + A\epsilon_1^\star + (1 - A)\epsilon_0^\star \ . \end{aligned}$$

Thus, if we define

$$U := A\epsilon_1^\star + (1 - A)\epsilon_0^\star \ ,$$

the coefficients in (3.12) are naturally identified as

$$\begin{aligned} \beta_0 &= \delta_0^\star \\ \beta_1 &= \delta_1^\star - \delta_0^\star \\ \beta_2 &= \gamma_0^\star \\ \beta_3 &= \gamma_1^\star - \gamma_0^\star \ . \end{aligned}$$

Moreover, because $\gamma_1^\star$ and $\gamma_0^\star$ are the arm-specific population least-squares slopes, the residuals satisfy the corresponding orthogonality conditions within each treatment arm. This implies that U is orthogonal to 1, A, W, and AW, so the decomposition above is exactly the population linear projection of Y on $(1, A, W, AW)$.

Finally, since $E[W] = 0$, we have

$$\delta_1^\star = E[Y(1)] \ , \qquad \delta_0^\star = E[Y(0)] \ ,$$

and therefore

$$\beta_1 = \delta_1^\star - \delta_0^\star = \theta \ .$$

3.5 Concluding Remarks

These notes are based on past notes for Econ 480-3 that I have written for previous editions of this class. Other sources that contain useful related concepts include the books by Hansen [2022], and the one by Angrist and Pischke [2008]. The results in Angrist [1998] are related to recent results on treatment effects with delayed outcomes, Bugni et al. [2026], and contamination bias when A is not binary, Goldsmith-Pinkham et al. [2022].

3.6 Problems

Problem 3.1 *Prove* (3.5)

Problem 3.2 *Complete all of the steps required to go from (3.4) to (3.8)*

Problem 3.3 *Consider the characterization of β_A in (3.8). Show that*

$$\beta_A = \sum_{w \in \mathcal{W}} \omega(w)\Delta(w) \ , \tag{3.13}$$

where the weights $\omega(w)$ are non-random and satisfy $\omega(w) \geq 0$ for all $w \in \mathcal{W}$ and $\sum_{w \in \mathcal{W}} \omega(w) = 1$.

Problem 3.4 *Using Angrist(1998)'s dataset, do the following regressions separately for the whites and non-whites:*

(a) regression of earnings on the veteran status only, i.e., the difference in means estimator;

(b) regression saturated in discrete covariates as in (3.2), only including the year of birth as the covariate;

(c) regression fully-saturated in discrete covariates as in (3.3), only including the year of birth as the covariate;

Report and compare your estimates for β_A with column (2) and (4) in Table 3.1. The variables in the dataset you may use are the following:

- *DNWHITE: dummy variable, 0 if white, 1 otherwise.*
- *DVET: dummy variable, 1 if veteran, 0 otherwise.*
- *DOBYY: year of birth.*
- *EARNVAR: earnings.*
- *YEAR: the year corresponding to the earnings record. (To replicate Table 3.1, you may select the years 1988-1991.)*

Problem 3.5 *Consider the common-slope class*

$$\hat{\theta}_{\text{adj}}(\gamma) = \bar{Y}_1 - \bar{Y}_0 - (\bar{W}_1 - \bar{W}_0)'\gamma,$$

and let $\gamma^\dagger$ denote the oracle minimizer of the corresponding asymptotic variance criterion $V(\gamma)$. Assume $E[W] = 0$, and define

$$Q_a := \text{Var}(W \mid A = a), \qquad \Lambda_a := \text{Cov}(Y, W \mid A = a), \qquad a \in \{0, 1\}.$$

Also let β_W^{pool} be the population slope from the pooled linear projection of Y on $(1, W)$ - the intercept is not relevant here given that $E[W] = 0$.

(a) Show that

$$\gamma^\dagger = \left(\frac{Q_1}{\pi} + \frac{Q_0}{1-\pi}\right)^{-1} \left(\frac{\Lambda_1}{\pi} + \frac{\Lambda_0}{1-\pi}\right).$$

(b) Show that

$$\beta_W^{\text{pool}} = \left(\pi Q_1 + (1-\pi)Q_0\right)^{-1}\left(\pi\Lambda_1 + (1-\pi)\Lambda_0\right).$$

(c) Conclude that, in general, $\gamma^\dagger \neq \beta_W^{\text{pool}}$*. Comment.*

(d) What happens when $\pi = 0.5$*? Comment.*

Problem 3.6 *Show that* (3.11) *holds.*

Problem 3.7 *Prove Proposition 3.3.*

Bibliography

J. D. Angrist. Estimating the labor market impact of voluntary military service using social security data on military applicants. *Econometrica*, 66 (2):249–288, 1998.

J. D. Angrist and J.-S. Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2008.

F. A. Bugni, I. A. Canay, and S. McBride. Decomposition and interpretation of treatment effects in settings with delayed outcomes. *Journal of Econometrics*, 2026. doi: doi.org/10.1016/j.jeconom.2025.106160.

P. Goldsmith-Pinkham, P. Hull, and M. Kolesár. Contamination bias in linear regressions. Technical report, National Bureau of Economic Research, 2022.

B. Hansen. *Econometrics*. Princeton University Press, 2022.

4

Selection on Observables

In this chapter, we shift our attention to observational studies. These are settings where the treatment A is not randomly assigned but, rather, is chosen by the units (individuals, families, firms, etc.). The key implication of an observational study is that the treatment, A, is typically **not** independent of the potential outcomes $Y(a)$. We will explore why this is the case below.

For our discussion, we use the following notation. We observe a binary treatment A, a real-valued outcome of interest Y, and some additional *pretreatment* covariates, denoted by W, which take values in $\mathbf{R}^{d_w}$. Therefore, the observed variables are (Y, A, W). Note that the term *pretreatment covariates* emphasizes that W represents variables that are realized *before* the treatment is chosen or assigned.

Throughout the lecture, we will use the following notation:

$$\mu_a(w) := E[Y(a) \mid W = w] \tag{4.1}$$

to denote the conditional expectations of potential outcomes given W,

$$\sigma_a^2(w) := \mathrm{Var}[Y(a) \mid W = w] \tag{4.2}$$

to denote the conditional variances of potential outcomes given W, and

$$\pi(w) := E[A \mid W = w] = P\{A = 1 \mid W = w\} \tag{4.3}$$

to denote the so-called propensity score.

The goal of this and the next chapter is to estimate causal parameters of interest (in particular, the ATE) under the assumption that, conditional on W, A is "as-good-as-random" even if we do not control it. We will first review the assumptions and their consequences for identification, then examine three methods for estimating average treatment effects (ATE): matching, regression, and propensity score weighting.

4.1 Observational Studies and Selection Bias

Observational studies are studies where the researcher does not control the assignment of the treatment. The researcher observes the treatment and the

outcome and tries to infer the causal effect of the treatment on the outcome. One of the immediate consequences of dealing with observational data is that the important assumption of treatment exogeneity, i.e.,

$$A \perp\!\!\!\perp (Y(a) : a \in \mathcal{A}) \ ,$$

is usually quite difficult to defend.

The inability to control the assignment of the treatment does not mean that one cannot think about the following question: how would the study be conducted if it were possible to do it by controlled experimentation? Thought experiments are useful to properly characterize the counterfactual question of interest and to clarify the assumptions needed to answer it in terms of potential outcomes. Indeed, as we will see throughout this class, many of the ideas in causal inference with observational studies are deeply connected to those with randomized experiments.

Example 4.1 (Job Training Program) LaLonde (1986) was interested in the causal effect of a job training program on earnings. He compared the results based on a randomized experiment to the results based on observational studies. LaLonde (1986) found that many traditional econometric methods for observational studies gave quite different estimates compared to the estimates based on the experimental data. Dehejia and Wahba (1999) re-analyzed the data using methods motivated by causal inference, and found that those methods can recover the experimental gold standard. Since then, this became a canonical example in causal inference with observational studies. ■

As we discussed in Chapter 1.6, random assignment is rarely compelling with observational data where agents choose A typically maximizing some criterion function (utility, profits, etc.). At the time, we defined selection into the treatment state A if

$$Y(a)|A = a \text{ is distributed } \textbf{differently} \text{ from } Y(a)|A = a' \text{ for } a \neq a' \ .$$

Consider the following concrete examples.

Example 4.2 (Educational Attainment and Income) Consider the effect of pursuing higher education (e.g., attending college) on income. People who choose to attend college may differ systematically from those who do not, such as being more motivated or having more financial resources. This means that $Y(0)$ may be systematically higher for those choosing $A = 1$ relative to those choosing $A = 0$. ■

Example 4.3 (Health Interventions and Health Outcomes) Consider a study examining the effect of a particular medical treatment (e.g., a new drug or therapy) on patient recovery rates. In an observational setting, patients who choose or are selected to receive the treatment might have different characteristics from those who do not, such as being sicker. This means that $Y(0)$ may

be systematically lower for those choosing $A = 1$ relative to those choosing $A = 0$. ■

For the moment we will focus attention to mean effects, and so the parameter of interest could be the average treatment effect (ATE), $\theta := E[Y(1) - Y(0)]$, the average treatment effect on the treated (ATT),

$$\theta_{\text{t}} := E[Y(1) - Y(0) \mid A = 1] \ ,$$

or the average treatment effect on the untreated (ATU),

$$\theta_{\text{u}} := E[Y(1) - Y(0) \mid A = 0] \ .$$

Ignoring covariates for the time being and focusing on the ATT and the ATU, it follows immediately from $Y = Y(A)$ that

$$\begin{aligned} \theta_{\text{t}} &= E[Y(1) \mid A = 1] - E[Y(0) \mid A = 1] \\ &= E[Y \mid A = 1] - E[Y(0) \mid A = 1] \ , \end{aligned}$$

and

$$\begin{aligned} \theta_{\text{u}} &= E[Y(1) \mid A = 0] - E[Y(0) \mid A = 0] \\ &= E[Y(1) \mid A = 0] - E[Y \mid A = 0] \ . \end{aligned}$$

In the above two formulas of θ_{t} and θ_{u}, the quantities $E[Y \mid A = 1]$ and $E[Y \mid A = 0]$ are identified from the observed data, but the quantities $E[Y(0) \mid A = 1]$ and $E[Y(1) \mid A = 0]$ are not. The latter two are *counterfactuals* because they are the expected value of the potential outcomes corresponding to the treatment level that is the *opposite* of the actual received treatment.

We can then decompose the usual difference in means contrast in terms of the ATT or the ATU to obtain:

$$\begin{aligned} E[Y \mid A = 1] - E[Y \mid A = 0] &= \theta_{\text{t}} + E[Y(0) \mid A = 1] - E[Y(0) \mid A = 0] \\ &= \theta_{\text{u}} + E[Y(1) \mid A = 1] - E[Y(1) \mid A = 0] \ . \end{aligned}$$

This shows that the difference in means contrast is biased for either the ATT or the ATU in settings in which we cannot guarantee that $A \perp\!\!\!\perp (Y(1), Y(0))$. The bias terms measure the differences in the means of the potential outcomes across the treatment and control groups, which would be expected to be different when agents choose A with knowledge of the potential outcomes.

Example 4.4 Recall the application in Angrist (98) Angrist [1998] that we presented in Example 3.1. Column (2) in Table 3.1 shows the differences in means for veteran and non-veterans earnings (among applicants to the US army). If we believe the US military tends to be picky about its soldiers and takes only high school graduates with test scores in the upper half of the test score distribution, then it follows that A would not be independent of potential

outcomes. That is, the resulting positive screening would generate positive selection bias in naive comparisons of veteran and non-veteran earnings, i.e.,

$$E[Y(0) \mid A = 1] - E[Y(0) \mid A = 0] > 0 \ .$$

Indeed, the results in Column (2) of Table 3.1 shows that veterans, on average, earn more than non-veterans. ■

4.2 Selection on Observables

Perhaps the simplest and most immediate relaxation of random assignment is to assume that the assignment of the treatment is independent of the potential outcomes conditional on the observed covariates W. This is known as selection on observables, conditional independence, strong ignorability, or unconfoundedness. We formally define it as follows,

Assumption 4.1 (unconfoundedness)

$$(Y(a) : a \in \mathcal{A}) \perp\!\!\!\perp A \mid W \ . \tag{4.4}$$

The assumption says that conditional on W, treatment is as-good-as randomly assigned. This means that there could be selection into the treatment state A, but that the entire selection is driven by the treatment (or action) to be correlated with W, which in turn may be correlated with the potential outcomes. There is an alternative version of this assumption that only requires $Y(a) \perp\!\!\!\perp A \mid W$ for all $a \in \mathcal{A}$, but for all practical purposes the distinction between these two is immaterial and so we work with (4.4).

The assumption in (4.4) implies that

$$E[Y(0) \mid A = 1, W] - E[Y(0) \mid A = 0, W] = 0$$

and

$$E[Y(1) \mid A = 1, W] - E[Y(1) \mid A = 0, W] = 0 \ .$$

In other words, the differences in the means of the potential outcomes across the treatment and control groups are entirely due to the difference in the observed covariates W. So given the same value of the covariates, the potential outcomes have the same means across the treatment and control groups.

One may wonder what type of models would satisfy this assumption. Heuristically, the assumption rules out the possibility of unobserved factors that affect both the treatment A and the outcomes Y simultaneously. Those "common shocks" or "common factors" are known as *confounders*, which ex-

plains the terminology unconfoundedness for Assumption 4.1. In terms of actual data generating processes (DGP), suppose that

$$\begin{aligned} Y(0) &= f_0(W, \epsilon_0, V) \\ Y(1) &= f_1(W, \epsilon_1, V) \\ A &= I\{g(W, \nu, V) \geq 0\} \ , \end{aligned}$$

for some functions (f_0, f_1, g) and unobservables $(\epsilon_0, \epsilon_1, \nu, V)$ that satisfy $(\epsilon_0, \epsilon_1) \perp\!\!\!\perp \nu$. If we ignore the presence of V for a second, then it is intuitive that such a DGP would satisfy the condition in (4.4) since, conditional on W, A would only be a function of ν and potential outcomes would only be a function of ϵ_a, which are independent by assumption. However, the presence of V breaks the argument and creates a model where, even conditional on W, A and $Y(a)$ are correlated due to their common unobservable V. Therefore, whenever we invoke an assumption like the one in (4.4) we are implicitly assuming that common unobserved factors like V are not present - or, in other words, that any common factor between A and Y are captured in the observed covariates W.

The assumption in (4.4) not only delivers identification of the conditional expectations of potential outcomes, as we show below, but it is also identifies the conditional distributions of the potential outcomes. To see this, consider the following arguments,

$$\begin{aligned} P\{Y \leq y \mid A = a, W = w\} &= P\{Y(a) \leq y \mid A = a, W = w\} \\ &= \frac{P\{Y(a) \leq y, A = a \mid W = w\}}{P\{A = a \mid W = w\}} \\ &= \frac{P\{Y(a) \leq y \mid W = w\} P\{A = a \mid W = w\}}{P\{A = a \mid W = w\}} \\ &= P\{Y(a) \leq y \mid W = w\} \\ &:= F_a(y|w) \ , \end{aligned}$$

where the first equality follows from $Y = Y(A)$, the second equality implicitly assumed that $P\{A = a \mid W = w\} \neq 0$, and the third equality invoked (4.4). The condition to avoid zero denominators is known as *overlap*, and in the case where A is binary is written in terms of the propensity score as follows,

$$0 < \pi(W) < 1 \text{ a.s.} \tag{4.5}$$

We can then conclude that under (4.4) and (4.5), the conditional distributions of the potential outcomes are identified by the observed data. The identification of $F_a(y|w)$ via $P\{Y \leq y \mid A = a, W = w\}$ also implies that

$$\mu_a(w) := E[Y(a) \mid W = w] = E[Y \mid A = a, W = w] \ . \tag{4.6}$$

That is, the conditional expectations of the potential outcomes are identified by the conditional expectations of the observed outcomes conditional on

(A, W). Due to this result, we will use $\mu_a(w)$ to denote either $E[Y(a) \mid W = w]$ or $E[Y \mid A = a, W = w]$ in what follows. Note that this immediately implies that the conditional average treatment effect

$$\theta(W) := E[Y(1) - Y(0) \mid W] = \mu_1(W) - \mu_0(W)$$

is identified by the conditional mean contrast $\Delta(W)$,

$$\Delta(W) := E[Y|A = 1, W] - E[Y|A = 0, W] \ .$$

By the LIE, this result automatically implies identification of the ATE under selection on observables,

$$\begin{aligned} \theta &= E[Y(1) - Y(0)] \\ &= E[\mu_1(W) - \mu_0(W)] \\ &= E[E[Y|A = 1, W] - E[Y|A = 0, W]] \ , \end{aligned} \tag{4.7}$$

where the second equality follows from the LIE, the third equality follows from (4.6). This last expression motivates the two alternative approaches to estimate θ that we discuss below: matching and regression.

Another implication of selection on observables is that it ensures that the conditional versions of the treatment effects we previously discussed, i.e.,

$$\theta(W) := E[Y(1) - Y(0) \mid W] \tag{4.8}$$

$$\theta_{\mathrm{t}}(W) := E[Y(1) - Y(0) \mid W, A = 1] \tag{4.9}$$

$$\theta_{\mathrm{u}}(W) := E[Y(1) - Y(0) \mid W, A = 0] \ , \tag{4.10}$$

are all the same,

$$\theta(W) = \theta_{\mathrm{t}}(W) = \theta_{\mathrm{u}}(W) \ .$$

This is not true for the unconditional counter-parts, since $\theta = E[\theta(W)]$ integrates over the distribution of W, $\theta_{\mathrm{t}} = E[\theta_{\mathrm{t}}(W) \mid A = 1]$ integrates using the conditional distribution $W|A = 1$, and $\theta_{\mathrm{u}} = E[\theta_{\mathrm{u}}(W) \mid A = 0]$ integrates using the conditional distribution $W|A = 0$. Since treatment is not randomly assigned, we expect all of these distributions to be different and so, in general,

$$\theta \neq \theta_{\mathrm{t}} \neq \theta_{\mathrm{u}} \ .$$

4.3 Estimation of the ATE

Let (Y, A, W) be a random vector where Y takes values in $\mathbf{R}$, A takes values in $\mathcal{A} = \{0, 1\}$, and W takes values in $\mathbf{R}^{d_w}$. We denote by P the distribution of (Y, A, W) and assume we have access to a random sample of size n from P that we denote by

$$\{(Y_i, A_i, W_i) : 1 \leq i \leq n\} \ .$$

We assume unconfoundedness as in (4.4) and overlap as in (4.5). Our goal is to estimate the ATE $\theta := E[Y(1) - Y(0)]$ or the ATT $\theta_{\mathrm{t}} := E[Y(1) - Y(0) \mid A = 1]$. We do so in three different ways, two of which we discuss today.

4.3.1 Matching

Matching estimators are better understood when the covariates W are discrete, taking values in a finite set $\mathcal{W}$. In this case, the identification of θ follows from (4.7) with a similar argument applying to θ_{t}. In order to define the so-called matching estimator of θ in this case, let

$$\hat{p}_n(w) := \frac{1}{n} \sum_{i=1}^{n} I\{W_i = w\}$$

denote the estimator of $p(w) := P\{W = w\}$ for all $w \in \mathcal{W}$, and let

$$n_{a,w} := \sum_{i=1}^{n} I\{A_i = a, W_i = w\}$$

denote the number of observations with $A = a$ and $W = w$. Note that $n_{a,w} > 0$ with probability approaching one under the overlap assumption and $p(w) > 0$. With this notation, the natural sample analog of θ is given by

$$\hat{\theta}_{n,\mathrm{mat}} := \sum_{w \in \mathcal{W}} \hat{p}_n(w)(\bar{Y}_{1,w} - \bar{Y}_{0,w}) \tag{4.11}$$

where

$$\bar{Y}_{a,w} := \frac{1}{n_{a,w}} \sum_{i=1}^{n} Y_i I\{A_i = a, W_i = w\} \ .$$

In words, $\hat{\theta}_{n,\mathrm{mat}}$ is a weighted average of the within cell $W = w$ differences in averages between units that are treated and untreated.

Example 4.5 Consider again the application in Angrist (98) Angrist [1998], where the covariates are discrete. Column (3) in Table 3.1 shows the matching estimates for the ATT, which only differ from (4.11) by the fact that $\hat{p}_n(w)$ is replaced by an estimator of $P\{W = w \mid A = 1\}$. Despite the fact that the matching and regression estimates control for the same variables, the estimand β_A in (3.2) is not equal to the ATT as it involves a set of weights that are not equal to $P\{W = w \mid A = 1\}$, see Problem A.3. In particular, matching uses the distribution of covariates among the treated to weight covariate-specific estimates into an estimate of the effect of treatment on the treated, while saturated-in-W regressions produce a variance-weighted average of these effects. See Problem 4.5 for more details. ■

The beauty and simplicity of matching estimators when W includes continuously distributed random variables quickly disappears. In this case, matching estimators impute the missing potential outcomes by using only the outcomes of nearest-neighbor units from the opposite treatment group. That is, the main idea is to find a "match" in the treatment group ($A = 1$) and control group ($A = 0$) with the same (or as close as possible) value of W, i.e., $W = w$. In that respect, matching is similar to nonparametric kernel regression methods, with the number of neighbors playing the role of the bandwidth in kernel regression. A formal difference is that the asymptotic distribution is derived conditional on the implicit bandwidth, that is, the number of neighbors, which is often fixed. In this class, however, we will not dwell on the details of the asymptotic approximations of matching estimators.

The matching estimator of θ can be formally defined as follows. For a fixed value q, let $j_q(i)$ be the index $j \in \{1, \ldots, n\}$ that solves

- Opposing treatment: $A_j = 1 - A_i$
- Opposing qth closest to i: $\sum_{s:A_s=1-A_i} I\{M_{is} \leq M_{ij}\} = q$,

where $M_{ij} = \|W_i - W_j\|$ is the distance (or matching metric) between units i and j in terms of the covariates. $j_q(i)$ is the index of the unit that is the qth closest to unit i in terms of the covariate values, among the units with the *treatment opposite* to that of unit i. Let $\mathcal{J}_q(i)$ denote the set of indices for the *first q matches* for unit i:

$$\mathcal{J}_q(i) := \{j_1(i), \ldots, j_q(i)\} .$$

The matching estimator of θ is given by

$$\hat{\theta}_{n,\text{mat}} := \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i(1) - \hat{Y}_i(0)) \quad \text{where} \quad \hat{Y}_i(a) := \begin{cases} Y_i & \text{if } A_i = a \\ \frac{1}{q}\sum_{j\in\mathcal{J}_q(i)} Y_j & \text{if } A_i \neq a \end{cases} .$$

Abadie and Imbens (2006) Abadie and Imbens [2006] study the asymptotic properties of $\hat{\theta}_{n,\text{mat}}$ under a fixed number of matches (as $n \to \infty$) and show that the estimator is consistent, but that its asymptotic normality and rates of convergence critically depend on the dimension of W (or, at least, to the dimension of the continuous elements of W). The bias of this estimator is of order $O(n^{-1/k_c})$, where k_c is the dimension of the (cont.) covariates, while the variance is of order $O(1/n)$. It follows that $\sqrt{n}\text{Bias} \to 0$, C, or ∞ if $k_c = 1$, $k_c = 2$, or $k_c > 2$, respectively. So, if $k_c > 2$, this estimator is not $\sqrt{n}$ asymptotically normal. The bootstrap is also known to be invalid, and other resampling tools we will discuss later on in class (like subsampling), are only valid for $k_c \leq 2$. Partly due to these reasons, matching is rarely viewed as the go-to method in settings with many covariates.

4.3.2 Regression

An alternative method to estimate θ under unconfoundedness and overlap is known as regression or imputation. We start with the characterization of θ in (4.7), which we can re-write as follows

$$\theta = \int (\mu_1(w) - \mu_0(w)) dF_W(w) \ , \tag{4.12}$$

where F_W here denotes the distribution function of W. It follows from there that we can construct a consistent estimator of θ in two steps. In the first step we estimate the conditional expectations $\mu_1(w)$ and $\mu_0(w)$ non-parametrically. If we denote these estimators by $\hat{\mu}_{n,a}(w)$ for $a \in \mathcal{A}$, the second step simply involves taking a sample average of the difference,

$$\hat{\theta}_{n,\text{reg}} := \frac{1}{n} \sum_{i=1}^{n} \left(\hat{\mu}_{n,1}\left(W_i\right) - \hat{\mu}_{n,0}\left(W_i\right) \right) \ . \tag{4.13}$$

We refer to this estimator as the "regression" estimator, but it is also usually referred to as the "imputation" estimator, since we essentially use the estimated functions $\hat{\mu}_{n,a}(w)$ to impute for the conditional means we do not get to observe. This estimator is consistent and asymptotically normal under certain conditions on the properties of the estimators $\hat{\mu}_{n,a}(w)$ for $a \in \mathcal{A}$ that we will discuss later in class.

It is important to note that the terminology "regression" does not necessarily mean "linear" regression. In fact, the estimator in (4.13) is most often based on non-parametric estimators of $\hat{\mu}_{n,a}(w)$, including essentially any of the estimators you learned in Econ 480-2 like kernel regression, local polynomials, sieves, series, lasso-type methods, neural nets, random forests, and others. In particular, this is *not* an estimator associated with the saturated-in-W regression in (3.2). When W is discrete, the approach in (4.13) would require separate regressions for $A = 1$ and $A = 0$ (which, in turn, would be equivalent to a fully saturated regression that includes interaction terms between A and I_w).

Relying on non-parametric estimators for $\mu_a(w)$ is certainly not the only way to approach this problem. It is not rare for researchers to instead rely on parametric models for $\mu_a(w)$, including a linear model of the form

$$\mu_a(w) = \delta_a + w'\gamma_a \ ,$$

where δ_a is a scalar and γ_a is a d_w-dimensional vector of parameters. This is equivalent to assuming a linear model for potential outcomes; i.e.,

$$Y(0) = \delta_0 + W'\gamma_0 + \epsilon_0 \tag{4.14}$$

$$Y(1) = \delta_1 + W'\gamma_1 + \epsilon_1 \tag{4.15}$$

where $E[\epsilon_a \mid W] = 0$. This model implies that treatment effect is given by

$$\Delta = Y(1) - Y(0) = \delta_1 - \delta_0 + W'(\gamma_1 - \gamma_0) + \epsilon_1 - \epsilon_0 \ ,$$

and so it is heterogeneous due to the presence of $\epsilon_1 - \epsilon_0$ (even after conditioning on W). Alternatively, the term $W'(\gamma_1 - \gamma_0)$ captures "observed" heterogeneity while the term $\epsilon_1 - \epsilon_0$ captures "unobserved" heterogeneity. It follows from this representation that

$$\theta(W) = \delta_1 - \delta_0 + W'(\gamma_1 - \gamma_0) \quad \text{and} \quad \theta = \delta_1 - \delta_0 + E[W]'(\gamma_1 - \gamma_0) \ .$$

We could then estimate θ by running two separate least squares regressions of Y on $(1, W)$ for each group (those units with $A_i = 1$ and those with $A_i = 0$). If we denote the least squares estimators of (δ_a, γ_a) by $(\hat{\delta}_{a,n}, \hat{\gamma}_{a,n})$, then

$$\hat{\theta}_{n,\text{ls}} = \hat{\delta}_{1,n} - \hat{\delta}_{0,n} + \left(\frac{1}{n} \sum_{i=1}^{n} W_i \right)' (\hat{\gamma}_{1,n} - \hat{\gamma}_{0,n}) \ . \tag{4.16}$$

This is equivalent to substituting the predicted values from each regression into (4.13). The same estimator can also be obtained in one step from a linear regression with interaction terms, but to make the coefficient on A equal to the estimator of the ATE, it is useful to recenter the covariates first. In particular, define

$$\tilde{W}_i := W_i - \bar{W}_n \quad \text{where} \quad \bar{W}_n := \frac{1}{n} \sum_{i=1}^{n} W_i \ .$$

If we now write the interacted regression using $\tilde{W}$ rather than W, then substituting the linear expressions for $Y(0)$ and $Y(1)$ yields

$$Y = \beta_0 + A\beta_1 + \tilde{W}'\beta_2 + A\tilde{W}'\beta_3 + U \ , \tag{4.17}$$

where

$$\begin{aligned} \beta_0 &= \delta_0 + \bar{W}_n'\gamma_0 \\ \beta_1 &= (\delta_1 - \delta_0) + \bar{W}_n'(\gamma_1 - \gamma_0) \\ \beta_2 &= \gamma_0 \\ \beta_3 &= \gamma_1 - \gamma_0 \\ U &= A\epsilon_1 + (1 - A)\epsilon_0 \ . \end{aligned}$$

Thus, after recentering the covariates, the coefficient on A is exactly

$$\beta_1 = (\delta_1 - \delta_0) + \bar{W}_n'(\gamma_1 - \gamma_0),$$

which is precisely the sample analogue of the ATE under the linear model. If we denote by $\hat{\beta}_{1,n}$ the least squares estimator of the coefficient on A from the interacted regression of Y on $(1, A, \tilde{W}, A\tilde{W})$, then

$$\hat{\beta}_{1,n} = \hat{\theta}_{n,\text{ls}} \ ,$$

where $\hat{\theta}_{n,\text{ls}}$ is the estimator in (4.13) obtained under the linear specification of

$\mu_a(w)$. This recentered interacted-regression representation is especially convenient in practice, since one can estimate the adjusted ATE and its standard error directly from a single regression. It is important to emphasize, however, that the use of linear regression to implement the estimator in (4.13) would only lead to a consistent estimator of θ provided the linear model is correctly specified (i.e., that the conditional means, or equivalently the potential outcomes, are indeed linear in W as in the previous model). More generally, the "regression" estimator of the ATE in (4.13) requires a good estimator of $\mu_a(w)$, whether that comes from a sufficiently accurate nonparametric estimator or from a correctly specified parametric model.

When W is discrete, as in the application of Example 3.1, then we could use a fully saturated regression in (A, I_w) to obtain a simple non-parametric estimator. In this case we can write the conditional expectations of potential outcomes as follows,

$$E[Y(0) \mid W = w] = E[Y \mid A = 0, W = w] = \sum_{w \in \mathcal{W}} \gamma_{0,w} I_w \tag{4.18}$$

$$E[Y(1) \mid W = w] = E[Y \mid A = 1, W = w] = \sum_{w \in \mathcal{W}} \gamma_{1,w} I_w \ . \tag{4.19}$$

It follows that in the case where W is discrete, we obtain

$$\theta = E[E[Y(1) - Y(0) \mid W]] = \sum_{w \in \mathcal{W}} (\gamma_{1,w} - \gamma_{0,w}) P\{W = w\} \ .$$

It is important to understand that this approach to identify, and later on estimate, θ or θ_{t} is different than the saturated-in-W regression in (3.2) we discussed in Example 3.1 and that is reported in Table 3.1. In that approach, the restriction $\gamma_{0,w} = \gamma_{1,w} = \gamma_w$ makes the approach not fully non-parametric and leads the coefficient on A, β_A, to become a weighted average of $\theta(W)$.

Finally, there is an alternative representation of the regression, or imputation, estimator that will become useful when we discuss estimators that are "doubly-robust". In particular, notice that the estimator in (4.13) imputes both conditional means, $\hat{\mu}_{n,1}(W_i)$ and $\hat{\mu}_{n,0}(W_i)$, for every observation. One can instead reduce the amount of imputation by using the observed outcome whenever it is available; for example, using Y_i when $A_i = 1$ instead of $\hat{\mu}_{n,1}(W_i)$, and using Y_i when $A_i = 0$ instead of $\hat{\mu}_{n,0}(W_i)$. This leads to the following variation of the "regression" estimator:

$$\tilde{\theta}_{n,\text{reg}} := \frac{1}{n} \sum_{i=1}^{n} A_i \left(Y_i - \hat{\mu}_{n,0} \left(W_i \right) \right) + \frac{1}{n} \sum_{i=1}^{n} (1 - A_i) \left(\hat{\mu}_{n,1} \left(W_i \right) - Y_i \right) \ . \tag{4.20}$$

For the purpose of our discussion today, the differences between $\hat{\theta}_{n,\text{reg}}$ and $\tilde{\theta}_{n,\text{reg}}$ are not that relevant and so we will revisit the discussion on the relative merits of these two variants later on in class.

4.4 Concluding Remarks

The material today borrows from several useful sources, including notes by Alex Torgovitsky, class notes by Wager [2020], and publicly available notes by Ding [2023]. I want to particularly thank Alex for sharing his source notes with me. In addition to these resources, the paper by Imbens [2004] provides a good review of many of the concepts we covered today.

4.5 Problems

Problem 4.1 *Consider the application in Example 3.1 and assume that certain screening factors like fitness and interpersonal skills are relevant for the decision of being admitted into the military (and thus, for veteran status). Explain how would this be a concern for the interpretation of the matching estimates in Table 3.1 (i.e., their interpretation as ATT).*

Problem 4.2 *Assume unconfoundedness as in (4.4) and overlap as in (4.5). Show that*

$$\theta_{\mathrm{t}} := E[Y|A=1] - E[\mu_0(W) \mid A=1] \ ,$$

using arguments similar to those in (4.7).

Problem 4.3 *Let $\hat{\mu}_0(w)$ be a consistent estimator of $\mu_0(w)$ for each $w \in \mathcal{W}$. What causal parameter would the following imputation estimator be estimating?*

$$\frac{1}{n_1} \sum_{i \in \mathcal{I}_1} \left(Y_i - \hat{\mu}_{n,0}\left(W_i\right) \right)$$

where

$$n_a := \sum_{i=1}^{n} I\{A=a\} \quad \text{and} \quad \mathcal{I}_a := \{i : A_i = a\} \ .$$

Use your intuition, do not do any derivations.

Problem 4.4 *Assume unconfoundedness as in (4.4) and overlap as in (4.5). Assume that the covariates W are discrete and take values in $\mathcal{W}$. Show that the ATT $\theta_{\mathrm{t}} := E[Y(1) - Y(0) \mid A=1]$ is identified by*

$$\frac{\sum_{w \in \mathcal{W}} \pi(w) P\{W=w\} \Delta(w)}{\sum_{w' \in \mathcal{W}} \pi(w') P\{W=w'\}}$$

where $\Delta(w) := E[Y \mid A=1, W=w] - E[Y \mid A=0, W=w]$.

Problem 4.5 *Consider the characterization of β_A in Problem A.3 and the expression of the ATT you derived in Problem 4.4. Show that both estimators admit the representation*

$$\sum_{w \in \mathcal{W}} \omega(w)\Delta(w) \ , \tag{4.21}$$

where the weights $\omega(w)$ are non-random and satisfy $\omega(w) \geq 0$ for all $w \in \mathcal{W}$ and $\sum_{w \in \mathcal{W}} \omega(w) = 1$. Compare the weights of each of these estimands and describe conditions under which the two would be the same.

Problem 4.6 *To estimate the ATT, the relevant unknown is $\gamma := E[Y(0)|A = 1]$. Suppose that there are two covariates, X and Z. We are willing to assume that $Y(0) \perp A|X, Z$. If we were able to observe both X and Z, we know that we could identify γ using the usual selection on observables argument. However, suppose that we only observe X and that we try to use the same argument with only X. Denote the naive estimand of γ as*

$$\gamma_n = E[Y \mid A = 0].$$

Denote the selection on observables estimand of γ that only uses X as

$$\gamma_s = E[E[Y \mid A = 0, X] \mid A = 1].$$

Suppose for simplicity that X is binary. Let $\mu_{ax} = E[Y(0)|A = a, X = x]$ and $q_d = P[X = 1 \mid D = d]$ for $d \in \{0, 1\}$. Then by LIE, we can write

$$\begin{aligned}
\gamma &= \mu_{10}\,(1 - q_1) + \mu_{11} q_1 \\
\gamma_n &= \mu_{00}\,(1 - q_0) + \mu_{01} q_0 \\
\gamma_s &= \mu_{00}\,(1 - q_1) + \mu_{01} q_1
\end{aligned}$$

Claim that it is possible that $|\gamma_n - \gamma| < |\gamma_s - \gamma|$, so that the naive estimand is closer to the actual ATT than the estimand that matches on only X. Explain in words the assumptions you need to impose to get this result.

Problem 4.7 *Assuming the conditional variances of the potential outcomes are the same for $a = 0 = 1$ and $a = 0$, i.e., $\sigma_0^2(w) = \sigma_1^2(w) = \sigma^2(w)$, derive the asymptotic variance of the matching estimator defined in (4.11) when W is discrete. Specifically, show that $\sqrt{n}(\hat{\theta}_{n,mat} - \theta) \xrightarrow{d} N(0, V_{mat})$, where*

$$V_{mat} = \mathrm{Var}\left[\theta\left(W\right)\right] + \mathbb{E}\left[\frac{\sigma^2\left(W\right)}{\pi\left(W\right)\left(1 - \pi\left(W\right)\right)}\right].$$

Problem 4.8 *Suppose that the covariates W are discrete and take values in a finite set $\mathcal{W}$. Consider the regression/imputation estimator*

$$\hat{\theta}_{n,\mathrm{reg}} := \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}_{n,1}\left(W_i\right) - \hat{\mu}_{n,0}\left(W_i\right)\right)$$

where $\hat{\mu}_{n,a}(w)$ is obtained by sample averages within cells:

$$\hat{\mu}_{n,a}(w) = \bar{Y}_{a,w} \ .$$

Show that in this case

$$\hat{\theta}_{n,\text{reg}} = \sum_{w \in \mathcal{W}} \hat{p}_n(w)(\bar{Y}_{1,w} - \bar{Y}_{0,w}) \ .$$

Conclude that when W is discrete, the regression/imputation estimator coincides with the matching estimator in (4.11).

Bibliography

A. Abadie and G. W. Imbens. Large sample properties of matching estimators for average treatment effects. *econometrica*, 74(1):235–267, 2006.

J. D. Angrist. Estimating the labor market impact of voluntary military service using social security data on military applicants. *Econometrica*, 66 (2):249–288, 1998.

J. D. Angrist and J.-S. Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2008.

P. Ding. A first course in causal inference. arXiv:2305.18793, 2023.

G. W. Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, 86(1):4–29, 2004.

S. Wager. Causal inference. Standford University, 2020.

5

Selection on Observables II

Today we continue the analysis of an observational study where we observe a binary treatment A, a real-valued outcome of interest Y, and some additional pretreatment covariates that we denote by W and that take values in $\mathbf{R}^{d_w}$; so that the observed data are (Y, A, W). Throughout the lecture, we will use the notation,

$$\mu_a(w) := E[Y(a) \mid W = w] \quad \text{and} \quad \sigma_a^2(w) := \text{Var}[Y(a) \mid W = w] \tag{5.1}$$

to denote the conditional expectations and variances of potential outcomes given W, and

$$\pi(w) := E[A \mid W = w] = P\{A = 1 \mid W = w\} \tag{5.2}$$

to denote the *propensity score*. The main identifying assumption continues to be selection on observables, as defined in Assumption 4.1, which states that

$$(Y(a) : a \in \mathcal{A}) \perp\!\!\!\perp A \mid W \ . \tag{5.3}$$

5.1 The Role of the Propensity Score

An important result building on the selection on observables assumption shows that one need not condition simultaneously on *all* covariates. The following result shows that all biases due to observable covariates can be removed by conditioning solely on the *propensity score*. That is, the assumption in (5.3) implies the following assumption,

$$(Y(a) : a \in \mathcal{A}) \perp\!\!\!\perp A \mid \pi(W) \ . \tag{5.4}$$

To see this result, note that

$$\begin{aligned} P\{A=1 \mid Y(0),Y(1),\pi(W)\} &= E\Big[E[A \mid Y(0),Y(1),\pi(W),W] \mid Y(0),Y(1),\pi(W)\Big] \\ &= E\Big[E[A \mid Y(0),Y(1),W] \mid Y(0),Y(1),\pi(W)\Big] \\ &= E\Big[E[A \mid W] \mid Y(0),Y(1),\pi(W)\Big] \\ &= E\Big[\pi(W) \mid Y(0),Y(1),\pi(W)\Big] \\ &= \pi(W) \ . \end{aligned}$$

The first equality follows from the LIE, the second equality follows from the properties of conditional expectations (i.e., $E[X|Z,g(Z)] = E[X|Z]$), the third equality follows from selection on observables, the fourth equality by the definition of the propensity score, and the last equality from the properties of conditional expectations (i.e., $E[X|X] = X$). The same steps show that $P\{A=1 \mid \pi(W)\} = \pi(W)$, and so

$$P\{A=1 \mid Y(0),Y(1),\pi(W)\} = P\{A=1 \mid \pi(W)\}$$

and the result then follows. This result is due to Rosenbaum and Rubin (1983) Rosenbaum and Rubin [1983] and shows that the propensity score is a sufficient statistic for the treatment assignment mechanism, in the sense that it contains all the information in W that is relevant for the treatment assignment mechanism. The original covariates W can be general and have many dimensions, but the propensity score $\pi(W)$ is a one-dimensional scalar variable bounded between 0 and 1. Therefore, the propensity score reduces the dimension of the original covariates but still maintains the "ignorability".

The result we just proved has two implications. The first one is that we could use the propensity score as a dimension reduction technique, where we could condition on $\pi(W)$ instead of on W and move the dimension of the conditioning set from d_w to just 1. The second one is less immediate and essentially leads to alternative characterizations of the ATE and ATT parameters that are based on *weighting* instead of matching. We discuss these two implications below.

5.1.1 Propensity Score Stratification

The implication of (5.4) is that if we can partition the observations into groups with the same values of the propensity score $\pi(W)$, then we can identify and consistently estimate parameters like the ATE and ATT using similar arguments to those we used in the previous class. Of course there are two complications with this intuition. First, the propensity score is generally unknown and needs to be estimated from the observed data. Second, when W contains continuously distributed covariates, the propensity score may take a continuum of values in $[0,1]$. We start by ignoring these two challenges and directly

assuming that the propensity score is known and takes only K possible values $\{\pi_1, \ldots, \pi_K\}$. The second feature may arise, for example, when W is discrete and so $\pi_k := P\{A = 1 \mid W = w_k\}$ for all $w_k \in \mathcal{W}$.

The identifying assumption in this case reduces to

$$(Y(a) : a \in \mathcal{A}) \perp\!\!\!\perp A \mid \pi(W) = \pi_k \quad \text{for} \quad k = 1, \ldots, K \ .$$

Therefore, we have a stratified randomized experiment, with K independent experiments within strata of the propensity score. We can then identify θ using the following argument,

$$\begin{aligned}
\theta &= E[Y(1) - Y(0)] \\
&= E[E[Y(1) - Y(0) \mid \pi(W)]] \\
&= E[E[Y \mid A = 1, \pi(W)] - E[Y \mid A = 0, \pi(W)]] \\
&= \sum_{k=1}^{K} (E[Y|A = 1, \pi(W) = \pi_k] - E[Y|A = 0, \pi(W) = \pi_k]) P\{\pi(W) = \pi_k\}
\end{aligned}$$

which parallels the one we used in (4.7) under (5.3). This result shows that the ATE can be identified by a weighted average of the differences in the conditional expectations of the potential outcomes within strata defined by the propensity score, where the weights are given by $P\{\pi(W) = \pi_k\}$. When W is discrete, these weights are simply $P\{W = w\}$ for all $w \in \mathcal{W}$ (assuming $\pi(w) \neq \pi(w')$ for $w \neq w'$).

We can easily define the natural sample analog associated with propensity score stratification after introducing some additional notation. Let $\hat{p}_{n,k} := \frac{1}{n} \sum_{i=1}^{n} I\{\pi(W_i) = \pi_k\}$ denote the estimator of $P\{\pi(W) = \pi_k\}$ for all $k \leq K$, and let

$$n_{a,k} := \sum_{i=1}^{n} I\{A_i = a, \pi(W_i) = \pi_k\}$$

denote the number of observations with $A = a$ in the kth stratum. With this notation, we obtain

$$\hat{\theta}_{n,\text{pss}} := \sum_{k=1}^{K} \hat{p}_{n,k} (\bar{Y}_{1,k} - \bar{Y}_{0,k}) \tag{5.5}$$

where

$$\bar{Y}_{a,k} := \frac{1}{n_{a,k}} \sum_{i=1}^{n} Y_i I\{A_i = a, \pi(W_i) = \pi_k\} \ .$$

In words, $\hat{\theta}_{n,\text{pss}}$ is a weighted average of the within stratum differences in averages between units that are treated and untreated.

In the particular case where W is discrete, both of the difficulties we previously mentioned can be simultaneously dealt with. First, we can easily estimate $\{\pi(w) : w \in \mathcal{W}\}$ non-parametrically by

$$\hat{\pi}_n(w) := \frac{\sum_{i=1}^{n} I\{A_i = 1, W_i = w\}}{\sum_{i=1}^{n} I\{W_i = w\}} \quad \forall w \in \mathcal{W} \ . \tag{5.6}$$

Second, by virtue of W being discrete, presumably there are sufficient observations for each value of W to have both treated and untreated units for each stratum. It follows that $\hat{\theta}_{n,\text{pss}}$ would be defined in the same way as before, with $\hat{\pi}_n(w)$ replacing $\pi(w)$.

In general, the propensity score is not known and W is not discrete. In this case researchers often fit a statistical prediction model for $P\{A = 1 \mid W\}$ (for example, a logistic model or a non-parametric model) to obtain the estimated propensity score $\hat{\pi}_n(W)$. This estimated propensity score can take up to as many values as the sample size, but we can discretize it to approximate the simple case above. For example, we can discretize the estimated propensity score by its K quantiles, or sort the values of $\hat{\pi}_n(W_i)$ and then split the sample into K evenly sized strata using the sorted propensity score. Once we obtain the K strata, we proceed in the same way we previously did to obtain $\hat{\theta}_{n,\text{pss}}$. This makes the final estimator dependent on the model specification for the propensity score, so using an incorrect model for $\pi(W)$ would lead to bias in $\hat{\theta}_{n,\text{pss}}$. An important practical question is how to choose K. If K is too small, it is expected that selection on observables would not appropriately hold. If K is too large, then we may not have enough units within each stratum of the estimated propensity score and many strata would only have treated or control units. Therefore, we face a trade-off in practice. A good data dependent rule to choose K has not yet been established.

Example 5.1 Chan et al. [2016] used a subsample of the data from NHANES 2007–2008 to study whether participation in school meal programs leads to an increase in BMI for schoolchildren. The dataset has the following important covariates:

age	age
ChildSex	gender (1: Male, 0: Female)
black	race (1: Black, 0: otherwise)
mexam	ethnicity (1: Hispanic, 0 otherwise)
pir200 plus	Family above 200% of the federal poverty level
WIC	Participation in the special supplemental nutrition program
Food Stamp	Participation in food stamp program
fsdchbi	Childhood food security
AnyIns	Any insurance
RefSex	Gender of the adult respondent (1: Male, 0: Female)
RefAge	Age of the adult respondent

Using a logistic regression of the treatment status variable (participation in the school meal programs) on these covariates, Figure 5.1 shows the histograms of the estimated propensity scores with different numbers of strata.

Based on the estimated propensity scores, we can partition the observations into strata for different choices of $K \in \{5, 10, 20, 50\}$ and calculate the estimators $\hat{\theta}_{n,\text{pss}}$. The results are presented in Table 5.1

K	5	10	20	50
$\hat{\theta}_{n,\text{pss}}$	-0.115	-0.177	-0.206	-0.267

TABLE 5.1: $\hat{\theta}_{n,\text{pss}}$ for different strata

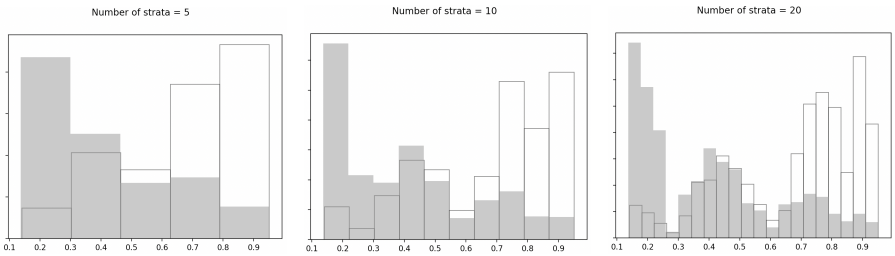


FIGURE 5.1: Histograms of the estimated propensity scores based on the NHANES data: white for the treatment group and gray for the control group

The estimates in Table 5.1 suggest that participation in school meal programs led to a lower BMI for students, while the naive difference in means estimator is positive (0.534). ■

5.1.2 Inverse Probability Weighting

The propensity score can be used not only as a tool for dimension reduction (since W is of dimension d_w while $\pi(W)$ is just a scalar). In fact, the propensity score leads to an alternative characterization of the ATE and ATT that exhibits certain benefits relative to the one in (4.7). To see this, consider the following argument,

$$\begin{aligned} E\left[\frac{YA}{\pi(W)}\right] &= E\left[\frac{1}{\pi(W)}E\left[Y(1)A \mid W\right]\right] \\ &= E\left[\frac{1}{\pi(W)}E[A \mid W]E[Y(1) \mid W]\right] \\ &= E[Y(1)] \ , \end{aligned}$$

where the first equality follows from the LIE, and the second equality follows from the selection on observables assumption in (4.4), and the last equality follows from the definition of the propensity score. Similarly, we can show that

$$E\left[\frac{Y(1-A)}{1-\pi(W)}\right] = E[Y(0)] \ .$$

Combining these two results, we obtain the following alternative characterization of θ,

$$\theta = E\left[\frac{YA}{\pi(W)}\right] - E\left[\frac{Y(1-A)}{1-\pi(W)}\right] \ . \tag{5.7}$$

This characterization is known as the inverse probability weighting (IPW) characterization of the ATE and shows that the ATE can be identified by a weighted average of the observed outcomes, where the weights are inversely proportional to the propensity score. Since the propensity score is in the denominator, this alternative characterization explicitly requires the same *overlap* assumption we previously invoked when we proved identification of $F_a(y)$ under selection on observables. We re-state the assumption here,

$$0 < \pi(W) < 1 \quad \text{a.s.} \tag{5.8}$$

Looking at the representation in (5.7), we can see immediately that in order to exploit this representation we would need to either know the propensity score $\pi(w)$ or be able to estimate it consistently. Denote by $\hat{\pi}_n(w)$ a consistent estimator of $\pi(w)$ and define,

$$\hat{\theta}_{n,\text{ipw}} := \frac{1}{n}\sum_{i=1}^{n}\left(\frac{A_i Y_i}{\hat{\pi}_n(W_i)} - \frac{(1-A_i)\,Y_i}{1-\hat{\pi}_n(W_i)}\right) . \tag{5.9}$$

We note again that in experiments where treatment is randomly assigned conditional on covariates (e.g., strata), the propensity score may be reasonably assumed to be known. In other settings that would not be the case, and so the consistency of $\hat{\theta}_{n,\text{ipw}}$ would fundamentally depend on whether $\pi(w)$ can be estimated consistently or not.

One somewhat unappealing feature of the estimator above is that the weights do not necessarily add to 1. This implies that this estimator is not invariant to location shifts on the outcomes (i.e., adding a constant to Y; see Problem 5.1). Specifically, the weights for the treated units add up to

$$\frac{1}{n}\sum_{i=1}^{n}\frac{A_i}{\hat{\pi}_n(W_i)} ,$$

which is equal to 1 in expectation but not in finite samples. There are variations of the estimator $\hat{\theta}_{n,\text{ipw}}$ that ensure that the weights add up to 1 by simply renormalizing the weights (dividing them by their sum for each group), i.e.,

$$\tilde{\theta}_{n,\text{ipw}} := \frac{\sum_{i=1}^{n}\frac{A_i Y_i}{\hat{\pi}_n(W_i)}}{\sum_{i=1}^{n}\frac{A_i}{\hat{\pi}_n(W_i)}} - \frac{\sum_{i=1}^{n}\frac{(1-A_i)Y_i}{1-\hat{\pi}_n(W_i)}}{\sum_{i=1}^{n}\frac{1-A_i}{1-\hat{\pi}_n(W_i)}}.$$

Remark 5.1 The estimator $\hat{\theta}_{n,\text{ipw}}$ is also known as the Horvitz-Thompson (HT) estimator, since it was proposed by Horvitz and Thompson [1952] in the context of survey sampling. In turn, the estimator $\tilde{\theta}_{n,\text{ipw}}$ is also known as the Hájek estimator due to Hájek [1971]. Extensive numerical evidence suggests that $\tilde{\theta}_{n,\text{ipw}}$ tends to perform better in finite samples. ■

It is interesting to contrast the regression estimator and the inverse-probability-weighting estimator of θ. The regression approach estimates the

two regression functions $\mu_a(w)$ nonparametrically and completely ignores the propensity score. The inverse probability weighting approach, on the other hand, estimates the propensity score nonparametrically and completely ignores the two regression functions. If appropriately implemented, both approaches lead to fully efficient estimators, but clearly their finite-sample properties may be very different, depending, for example, on the smoothness of the regression functions versus the smoothness of the propensity score. If there is only a single binary covariate, or more generally if there are only discrete covariates, the weighting approach with a fully nonparametric estimator for the propensity score is *numerically identical* to the regression approach with a fully nonparametric estimator for the two regression functions.

Example 5.2 Revisiting Example 5.1, we can obtain the IPW estimators based on the estimated propensity scores. Recall that $\pi(W)$ was estimated via a logit regression. Using the estimated IPW, the Horvitz-Thompson (HT) estimator - or original IPW estimator - is given by $\hat{\theta}_{n,\text{ipw}} = -1.517$, whereas the Hájek estimator - or re-normalized IPW estimator - is given by $\tilde{\theta}_{n,\text{ipw}} = -0.155$. ■

5.2 On the Asymptotic Efficiency of IPW Estimators

There is an interesting phenomenon that happens with IPW estimators that is also present in some other settings that involve weighting. It turns out that weighting observations by the inverse of the *true* (unknown) propensity score does not lead to efficient estimators, whereas weighting each observation by the inverse of a non-parametric estimate of the propensity score does lead to an efficient estimator. This point was formally derived by Hirano et al. [2003], and here we provide the basic intuition in the case where W is discrete.

We start by invoking the result in Problem 5.3, which allows us to write $\hat{\theta}_{n,\text{ipw}}$ as follows,

$$\hat{\theta}_{n,\text{ipw}} := \frac{1}{n}\sum_{i=1}^{n}\left(\frac{A_i Y_i}{\frac{n_{1,w}}{n_w}} - \frac{(1-A_i)\,Y_i}{1-\frac{n_{1,w}}{n_w}}\right) = \sum_{w\in\mathcal{W}}\frac{n_w}{n}(\bar{Y}_{1,w} - \bar{Y}_{0,w}) \ ,$$

where we used that $\hat{\pi}_n(w) := \frac{n_{1,w}}{n_w}$, with

$$n_{a,w} := \sum_{i=1}^{n} I\{A_i = a, W_i = w\}$$

$$n_w := \sum_{i=1}^{n} I\{W_i = w\}$$

and

$$\bar{Y}_{a,w} := \frac{1}{n_{a,w}} \sum_{i=1}^{n} Y_i I\{A_i = a, W_i = w\} \ .$$

Then consider a fixed value $w \in \mathcal{W}$ and let $p(w) := P\{W = w\}$ and $\theta(w) = E[Y(1) - Y(0) \mid W = w]$. We can expand the ipw estimator and write it as a function of terms we can easily analyze. Start by writing

$$\begin{aligned}
\hat{\theta}_{n,\text{ipw}} &= \sum_{w \in \mathcal{W}} \left(\frac{n_w}{n} \pm p(w) \right) (\bar{Y}_{1,w} - \bar{Y}_{0,w} \pm \theta(w)) \\
&= \theta + \sum_{w \in \mathcal{W}} p(w)(\bar{Y}_{1,w} - \bar{Y}_{0,w} - \theta(w)) + \sum_{w \in \mathcal{W}} \left(\frac{n_w}{n} - p(w) \right) \theta(w) \\
&+ \sum_{w \in \mathcal{W}} \left(\frac{n_w}{n} - p(w) \right) (\bar{Y}_{1,w} - \bar{Y}_{0,w} - \theta(w))
\end{aligned}$$

and then re-organize the terms to obtain

$$\sqrt{n}(\hat{\theta}_{n,\text{ipw}} - \theta) = R_{1,n} + R_{2,n} + R_{3,n} \tag{5.10}$$

where

$$\begin{aligned}
R_{1,n} &:= \sum_{w \in \mathcal{W}} \sqrt{\frac{n}{n_w}} p(w) \sqrt{n_w} (\bar{Y}_{1,w} - \bar{Y}_{0,w} - \theta(w)) \\
R_{2,n} &:= \sum_{w \in \mathcal{W}} \sqrt{n} \left(\frac{n_w}{n} - p(w) \right) \theta(w) \\
R_{3,n} &:= \sum_{w \in \mathcal{W}} \sqrt{n} \left(\frac{n_w}{n} - p(w) \right) (\bar{Y}_{1,w} - \bar{Y}_{0,w} - \theta(w)) \ .
\end{aligned}$$

The asymptotic behavior of $R_{1,n}$ follows from the joint convergence of $\{\sqrt{n_w} \left(\bar{Y}_{1,w} - \bar{Y}_{0,w} - \theta(w) \right) : w \in \mathcal{W}\}$, which in turn follows from some algebra and the CLT. We omit the details here, and simply state that such an exercise leads to

$$R_{1,n} \xrightarrow{d} N(0, V_1) \quad \text{where} \quad V_1 := E\left[\frac{\sigma_1^2(W)}{\pi(W)} + \frac{\sigma_0^2(W)}{1 - \pi(W)} \right] \ .$$

The asymptotic behavior of $R_{2,n}$ follows from the joint convergence of $\{\sqrt{n} \left(\frac{n_w}{n} - p(w) \right) : w \in \mathcal{W}\}$. which in turn directly follows from the CLT. We omit the details here, and simply state that such an exercise leads to

$$R_{2,n} \xrightarrow{d} N(0, V_2) \quad \text{where} \quad V_2 := V[\theta(W)] \ .$$

Finally, since $(\bar{Y}_{1,w} - \bar{Y}_{0,w} - \theta(w)) = o_p(1)$ for all $w \in \mathcal{W}$, $R_{3,n} = o_p(1)$ and we conclude that

$$\sqrt{n}(\hat{\theta}_{n,\text{ipw}} - \theta) \xrightarrow{d} N(0, V_1 + V_2) \ . \tag{5.11}$$

Remark 5.2 The result we just derived considers a setting where W is discrete with a finite support. Inference in this case may follow from standard arguments and the asymptotic approximation does not depend on the number of elements in $\mathcal{W}$. However, if W is continuous (or the cardinality of W is very large), this result would not provide an accurate asymptotic approximation because the sample size for each value $w \in \mathcal{W}$ would not be large enough to properly invoke the CLTs as we did before. Despite requiring different formal arguments, the conclusion in (5.11) also holds when W is continuous. ■

Remark 5.3 Next class we will learn that the limiting variance $V_1 + V_2$ is the so-called "semi-parametric efficiency bound" for estimating θ, which means that it is the best possible limiting variance that is achievable under certain assumptions. The IPW estimator is semiparametrically efficient not only when the regressors are discrete, but also when they are continuous under regularity conditions on $\pi(w)$ and $\mu_a(w)$. ■

An interesting phenomenon that arises in the analysis of the asymptotic variance of IPW estimators is that the estimator may lead to more accurate results when the propensity score is estimated than when the propensity score is *known*. That is, consider the following "oracle" version of the IPW estimator of θ,

$$\hat{\theta}^{\star}_{n,\text{ipw}} := \frac{1}{n}\sum_{i=1}^{n}\left(\frac{A_i Y_i}{\pi(W_i)} - \frac{(1-A_i)\,Y_i}{1-\pi(W_i)}\right) . \tag{5.12}$$

Problem 5.4 shows that this estimator is asymptotically normal, with asymptotic mean zero and variance given by

$$V^{\star} := V_1 + V_2 + V_3 , \tag{5.13}$$

where $V_3 > 0$. That is, the performance of the oracle IPW estimator is somewhat disappointing. Despite having access to the true propensity score $\pi(W)$, it underperforms relative to the "feasible" version. At a high level, the reason this phenomenon occurs is that the estimated propensity score corrects for local variability in the sampling distribution of the A_i (i.e., it accounts for the number of units that were *actually* treated in each group). Hirano et al. [2003] provide a detailed discussion of the asymptotics of IPW-style estimators; and in particular they discuss conditions under which IPW with non-parametrically estimated propensity scores can outperform oracle IPW.

5.3 Multivalued Treatments

The discussion so far has focused exclusively on the case where A is binary. One may then wonder to what extent these results translate to a context

where the treatment is multivalued; that is, $\mathcal{A} = \{0, 1, 2, 3, 4\}$ for example. In fact, many interesting counterfactual states are indeed multivalued. From the identification point of view, selection on observables naturally extends to a multivalued setting. Identification through matching or via non-parametric regression methods also extend well as long as the treatment continues to take discrete values. Propensity score weighting is more delicate. There are approaches that are based on "generalized" versions of the propensity score, but these approaches are not without subtleties as the propensity score no longer partitions the population into subgroups where we can obtain causal contrasts (at least not without proper averaging ex-post). The differences also affect how to interpret regression estimands where one runs a regression of the outcome on the treatment and covariates. For example, the result we derived in (3.8) where we showed that the LS slope was a weighted average of conditional average treatment effect (CATE) does not hold without additional assumptions when A is not binary; a feature that has most recently been discussed and formalized by Goldsmith-Pinkham et al. [2022]. In general, one should be careful when dealing with multivalued treatments or, equivalently, when there are multiple binary treatments, if the approach involves propensity score weighting or linear regressions that are expected to be interpreted as weighted averages of interesting causal effects.

5.4 Scope of Selection on Observables

Selection on observables continues to be a popular assumption in many social sciences and, more recently, in industry applications. But it is nevertheless an assumption that is difficult to digest in most economic applications, as inherent unobservables (preferences, private info, expectations) tend to play a role in how people make decisions and these, in turn, tend to depend on potential outcomes. That is, the idea that observationally identical people behave differently due to a coin flip is difficult to defend. One idea to feel more reassured about this approach to identification of causal effects is to "control for more"; that is, include a larger set of covariates. However, it can be shown (and we show this in Problem 4.5) that controlling for more covariates may increase the bias of estimators relative to using a subset of them. There is also an existing tension with overlap. That is, if we could perfectly explain A with W then $P\{A = 1|W\}$ would be either 0 or 1 and we wouldn't have the required variation. Better methods for choosing observables will not solve these problems, so the scope for machine learning techniques to provide guarantees about identification is limited. However, taking the identifying assumption as given and focusing instead on estimation accuracy, it is indeed the case that more modern ML techniques could lead to better estimators than the ones we discussed today. We'll cover one of these improvements next class.

5.5 Concluding Remarks

The material today borrows from several useful sources, including notes by Alex Torgovitsky, class notes by Stefan Wager Wager [2020], publicly available notes by Peng Ding Ding [2023], and the book by Angrist and Pischke [2008]. I want to particularly thank Alex for sharing his source notes with me. In addition to these resources, the paper by Guido Imbens Imbens [2004] provides a good review of many of the concepts we covered today.

5.6 Problems

Problem 5.1 *Show that $\hat{\theta}_{n,\text{ipw}}$ is not invariant to location transformations of the outcome. That is, if we redefine Y as $Y+c$ for a constant c this would affect the magnitude of $\hat{\theta}_{n,\text{ipw}}$. Show that $\tilde{\theta}_{n,\text{ipw}}$ does not suffer from this problem.*

Problem 5.2 *Show that $R_{2,n} := \sum_{w\in\mathcal{W}} \sqrt{n}\left(\frac{n_w}{n} - p(w)\right)\theta(w)$ satisfies*

$$R_{2,n} \xrightarrow{d} N(0, V_2) \quad \textit{where} \quad V_2 := V[\theta(W)] \ .$$

Problem 5.3 *Assume that W is discrete and takes K values. Show that $\hat{\theta}_{n,\text{ipw}}$, $\hat{\theta}_{n,\text{pss}}$, and $\hat{\theta}_{n,\text{mat}}$ and $\hat{\theta}_{n,\text{reg}}$, where the unknown functions are estimated non-parametrically, are all numerically equivalent.*

Problem 5.4 *Show that the oracle IPW estimator is asymptotically normal, with asymptotic mean zero and variance given by*

$$V^\star := V_1 + V_2 + V_3,$$

where $V_1 := E\left[\frac{\sigma_1^2(W)}{\pi(W)} + \frac{\sigma_0^2(W)}{1-\pi(W)}\right]$, $V_2 := V[\theta(W)]$ and $V_3 > 0$.

Problem 5.5 *Show that the IPW representation in (5.7) can be extended to the ATT. In particular, prove that*

$$\theta_{\text{t}} = \frac{1}{P\{A=1\}} E\left[AY - \frac{\pi(W)(1-A)Y}{1-\pi(W)}\right] \ .$$

Explain briefly why the weights in this representation differ from the ones in the IPW representation of the ATE.

Problem 5.6 *Consider the Hájek estimator*

$$\tilde{\theta}_{n,\text{ipw}} := \frac{\sum_{i=1}^{n} \frac{A_i Y_i}{\hat{\pi}_n(W_i)}}{\sum_{i=1}^{n} \frac{A_i}{\hat{\pi}_n(W_i)}} - \frac{\sum_{i=1}^{n} \frac{(1-A_i) Y_i}{1-\hat{\pi}_n(W_i)}}{\sum_{i=1}^{n} \frac{1-A_i}{1-\hat{\pi}_n(W_i)}} \ .$$

Show that this estimator is invariant to location transformations of the outcome. That is, show that if Y_i is replaced by $Y_i + c$ for some constant c, then $\tilde{\theta}_{n,\text{ipw}}$ does not change.

Bibliography

J. D. Angrist and J.-S. Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2008.

K. C. G. Chan, S. C. P. Yam, and Z. Zhang. Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(3):673–700, 2016.

P. Ding. A first course in causal inference. arXiv:2305.18793, 2023.

P. Goldsmith-Pinkham, P. Hull, and M. Kolesár. Contamination bias in linear regressions. Technical report, National Bureau of Economic Research, 2022.

J. Hájek. Comment on an essay on the logical foundations of survey sampling. In *Foundations of Statistical Inference*, page 236. Holt, Rinehart and Winston, 1971.

K. Hirano, G. W. Imbens, and G. Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71 (4):1161–1189, 2003.

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.

G. W. Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, 86(1):4–29, 2004.

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.

S. Wager. Causal inference. Standford University, 2020.

6

Augmented IPW and Double Robustness

Today we continue the study of observational data with a binary treatment A, a real-valued outcome Y, and pretreatment covariates W. The observed data consist of a random sample of size n from the distribution of (Y, A, W). As before, for each $a \in \mathcal{A}$ and $w \in \mathbf{R}^{d_w}$, we write

$$\begin{aligned} \mu_a(w) &:= E[Y(a) \mid W = w] \\ \sigma_a^2(w) &:= \mathrm{Var}[Y(a) \mid W = w] \\ \pi(w) &:= E[A \mid W = w] = P\{A = 1 \mid W = w\} \ . \end{aligned}$$

6.1 Introduction

In the previous lecture, under selection on observables, we studied two main approaches to identification and estimation of the average treatment effect,

$$\theta := E[Y(1) - Y(0)] \ .$$

The first approach was based on outcome regression. Using

$$\theta = E[\mu_1(W) - \mu_0(W)] \ ,$$

we obtained the regression (or imputation) estimator

$$\hat{\theta}_{n,\mathrm{reg}} = \frac{1}{n} \sum_{i=1}^{n} (\hat{\mu}_{n,1}(W_i) - \hat{\mu}_{n,0}(W_i)) \ .$$

The second approach was based on inverse probability weighting. Using

$$\theta = E\left[\frac{YA}{\pi(W)}\right] - E\left[\frac{Y(1-A)}{1-\pi(W)}\right] \ ,$$

we obtained the IPW estimator

$$\hat{\theta}_{n,\mathrm{ipw}} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{A_i Y_i}{\hat{\pi}_n(W_i)} - \frac{(1-A_i)Y_i}{1-\hat{\pi}_n(W_i)} \right) \ .$$

In this lecture we study the large-sample behavior of these estimators and introduce a third estimator of θ that combines the ideas behind both of them.

6.2 Semi-parametric Efficiency

Before we introduce the new estimator, we review some results on the efficiency bound for estimators of the ATE and the ATT. Throughout our discussion we maintain the unconfoundedness assumption in Assumption 4.1, and the overlap assumption. However, in order to do asymptotics properly we need a strengthening of the overlap assumption as described below.

Assumption 6.1 (overlap)

$$\eta < \pi(W) < 1 - \eta \text{ a.s. for some } \eta > 0 \ . \tag{6.1}$$

The formal results also require some smoothness assumptions on $\mu_a(w)$ and $\pi(w)$ that we omit for the moment. Formally, Hahn (1998) Hahn [1998] shows that for any reasonably well behaved estimator (called *regular*) of θ, denoted by $\tilde{\theta}_n$, that satisfies

$$\sqrt{n}(\tilde{\theta}_n - \theta) \xrightarrow{d} N(0, V) \ ,$$

for some limiting variance V, it must be the case that

$$V \geq V^* := E\left[\frac{\sigma_1^2(W)}{\pi(W)} + \frac{\sigma_0^2(W)}{1 - \pi(W)}\right] + \text{Var}[\theta(W)] \ . \tag{6.2}$$

The right hand side of (6.2) is known as the semi-parametric efficiency bound and provides a benchmark for how well an estimator of θ can do. Importantly, knowledge of the propensity score does not affect this efficiency bound. Hahn also computes the semi-parametric efficiency bound for the ATT, θ_{t}, both with and without knowledge of the propensity score. In this case, knowledge of the propensity score leads to a lower efficiency bound. The intuition that knowledge of the propensity score affects the efficiency bound for the ATT, but not for the ATE, goes as follows. Both are weighted averages of the treatment effect conditional on the covariates, $\theta(W)$. For the ATE the weight is proportional to the density of the covariates, whereas for the ATT the weight is proportional to the product of the density of the covariates and the propensity score. Knowledge of the propensity score implies one does not need to estimate the weight function and thus improves precision.

An immediate question is whether the estimators $\hat{\theta}_{n,\text{reg}}$ and $\hat{\theta}_{n,\text{ipw}}$ are asymptotically normal with variance V^* or not. The answer depends on two main elements: (a) how much smoothness we are willing to assume on the unknown functions $\mu_a(w)$ and $\pi(w)$, and (b) how well the non-parametric estimators of these functions perform, including the choice of tuning parameters. Under sufficiently strong conditions on both fronts, it is known that $\hat{\theta}_{n,\text{reg}}$ and $\hat{\theta}_{n,\text{ipw}}$ are asymptotically equivalent and asymptotically normal with variance V^*. In other words, both estimators are asymptotically efficient. Classical references for these results include Hahn (1998) Hahn [1998], Imbens (2006)

Imbens et al. [2006], and Hirano et al. (2003) Hirano et al. [2003]. Listing all the regularity conditions needed for non-parametric estimators of these functions (such as series, sieves, or kernels) is not something we want to emphasize in this class. Still, it is useful to highlight how demanding some of these conditions can be.

For example, if $\hat{\mu}_{n,a}(w)$ and $\hat{\pi}_n(w)$ are $\sqrt{n}$-consistent and asymptotically normal, then one can show that $\hat{\theta}_{n,\text{reg}}$ and $\hat{\theta}_{n,\text{ipw}}$ are asymptotically efficient. However, achieving the parametric rate is typically unrealistic with non-parametric estimators, including modern machine learning methods. A milder requirement is that these nuisance estimators be uniformly consistent at rate $o_P(n^{-1/4})$ (or $O_P(n^{-1/4})$ in some cases). For example, if we focus on $\hat{\theta}_{n,\text{ipw}}$ and assume the propensity score is s times continuously differentiable in W, then a condition such as $s/d_w \geq 7$ would require substantial smoothness. If we have 5 covariates, this means that $\pi(w)$ would need to have at least 35 continuous derivatives. More generally, if we let $\mathcal{W} \subseteq \mathbf{R}^{d_w}$ denote the support of W and assume this support is compact, then we would need enough conditions on $\pi(w)$ and $\hat{\pi}_n(w)$ to guarantee

$$\sup_{w \in \mathcal{W}} |\hat{\pi}_n(w) - \pi(w)| = o_P(n^{-1/4}) \; . \tag{6.3}$$

We will discuss in more detail the role of conditions like this in the next section.

6.3 Augmented IPW

So far, we have learned that we can estimate θ via regression adjustments or via inverse probability weighting. It turns out that we can combine the two representations of θ to obtain yet another estimator of θ that, as we will show later, exhibits several benefits over both $\hat{\theta}_{n,\text{reg}}$ and $\hat{\theta}_{n,\text{ipw}}$. This alternative estimator, typically known as the Augmented IPW estimator, is due to Robins et al. [1994] and takes the form

$$\hat{\theta}_{n,\text{aug}} = \frac{1}{n} \sum_{i=1}^{n} \Big(\hat{\mu}_{n,1}(W_i) - \hat{\mu}_{n,0}(W_i) + A_i \frac{Y_i - \hat{\mu}_{n,1}(W_i)}{\hat{\pi}_n(W_i)} - (1 - A_i) \frac{Y_i - \hat{\mu}_{n,0}(W_i)}{1 - \hat{\pi}_n(W_i)} \Big) \; .$$

Intuitively, AIPW begins with the regression estimator and then corrects its bias by applying IPW to the residuals from the outcome regressions. This construction leads to the key property of *double robustness*: the estimator remains consistent if either the outcome regressions are correctly specified (so

μ_a is correct) or the propensity score model is correctly specified (so π is correct). More formally, double robustness means that the estimator is consistent if:

- either both $\hat{\mu}_{n,1}(w)$ and $\hat{\mu}_{n,0}(w)$ are consistent;
- or $\hat{\pi}_n(w)$ is consistent.

Importantly, it does not require *all* these estimators to be consistent. To see this in more detail, consider the following two cases.

Case 1: $\hat{\mu}_{n,a}(w)$ are consistent

In this case, suppose that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}_{n,1}\left(W_i\right)-\hat{\mu}_{n,0}\left(W_i\right)\right) &= \frac{1}{n}\sum_{i=1}^{n}\left(\mu_1\left(W_i\right)-\mu_0\left(W_i\right)\right)\\ &\xrightarrow{P} E\left[\mu_1\left(W_i\right)-\mu_0\left(W_i\right)\right]\end{aligned}$$

consistently estimates θ, and where the last convergence follows from the LLN. The second term has mean zero using the following argument:

$$\begin{aligned}E\left[A_i\frac{Y_i-\mu_1(W_i)}{\bar{\pi}(W_i)}\right] &= E\left[E\left[A_i\frac{Y_i-E\left[Y_i\mid W_i,A_i\right]}{\bar{\pi}(W_i)}\mid W_i,A_i\right]\right]\\ &= E\left[E\left[\frac{Y_i(1)-\mu_1(W_i)}{\bar{\pi}(W_i)}\mid W_i,A_i=1\right]P\{A_i=1\mid W_i\}\right]\\ &= E\left[\left(\mu_1(W_i)-\mu_1(W_i)\right)\frac{\pi(W_i)}{\bar{\pi}(W_i)}\right]=0\ ,\end{aligned}$$

where we used $\mu_1(W_i)=E[Y_i(1)\mid W_i,A_i=1]$ and where the last equality does not depend on whether $\hat{\pi}_n(w)$ is consistent for $\pi(w)$ or not. A similar argument shows that $E\left[(1-A_i)\frac{Y_i-\mu_0(W_i)}{1-\bar{\pi}(W_i)}\right]=0$. Intuitively, even if the propensity score weights $1/\bar{\pi}(W_i)$ and $1/\left(1-\bar{\pi}(W_i)\right)$ are incorrect, they are multiplied by a zero-mean noise and thus will not matter asymptotically.

Case 2: $\hat{\pi}_n(W_i)$ is consistent

Continuing with the same logic as before, assume that $\hat{\pi}_n(W_i)=\pi(W_i)$ and $\hat{\mu}_{n,a}(w)=\bar{\mu}_a(w)\neq\mu_a(w)$ are non-stochastic so that we can ignore $o_P(1)$ terms. Rewrite $\hat{\theta}_{n,\text{aug}}$ as

$$\begin{aligned}\hat{\theta}_{n,\text{aug}} = &\frac{1}{n}\sum_{i=1}^{n}\left(\frac{A_iY_i}{\pi(W_i)}-\frac{\left(1-A_i\right)Y_i}{1-\pi(W_i)}\right)\\ &+\frac{1}{n}\sum_{i=1}^{n}\left(\bar{\mu}_1(W_i)\left(1-\frac{A_i}{\pi(W_i)}\right)-\bar{\mu}_0(W_i)\left(1-\frac{1-A_i}{1-\pi(W_i)}\right)\right)\ .\end{aligned}$$

The first term is the same as $\hat{\theta}_{n,\text{ipw}}$ and therefore consistently estimates θ. The second part has mean zero regardless of whether $\bar{\mu}_a(W_i)$ for $a \in \{0,1\}$ is correct (i.e., equal to the true conditional mean) or not. To see this, note that

$$\begin{aligned} E\left[\bar{\mu}_1(W_i)\left(1-\frac{A_i}{\pi(W_i)}\right)\right] &= E\left[\bar{\mu}_1(W_i)E\left[\left(1-\frac{A_i}{\pi(W_i)}\right) \mid W_i\right]\right] \\ &= E\left[\bar{\mu}_1(W_i)\left[1-\frac{E[A_i \mid W_i]}{\pi(W_i)}\right]\right] = 0 \ , \end{aligned}$$

where the last equality follows from $\left[1-\frac{E[A_i|W_i]}{\pi(W_i)}\right] = 0$. Similarly,

$$E\left[\bar{\mu}_0(W_i)\left(1-\frac{1-A_i}{1-\pi(W_i)}\right)\right] = 0 \ .$$

Intuitively, even if the regression adjustments $\hat{\mu}_{n,a}(W_i)$ are inconsistent, they are multiplied by a zero-mean noise and thus the inconsistency would not matter asymptotically.

6.4 Semiparametric Efficiency of the AIPW Estimator

The AIPW estimator of θ is not only doubly robust; it is also asymptotically optimal in the sense introduced in Section 6.2. That is, provided we estimate $\mu_a(w)$ and $\pi(w)$ in a reasonably accurate way (and we will discuss specific conditions under which this holds in just a minute), one can show that $\hat{\theta}_{n,\text{aug}}$ is first order equivalent to the oracle AIPW estimator

$$\begin{aligned} \hat{\theta}^*_{n,\text{aug}} = \frac{1}{n}\sum_{i=1}^{n}\Big(&\mu_1(W_i)-\mu_0(W_i) \\ &+A_i\frac{Y_i-\mu_1(W_i)}{\pi(W_i)} - (1-A_i)\frac{Y_i-\mu_0(W_i)}{1-\pi(W_i)}\Big) \end{aligned} \tag{6.4}$$

meaning that

$$\sqrt{n}\left(\hat{\theta}_{n,\text{aug}} - \hat{\theta}^*_{n,\text{aug}}\right) \xrightarrow{P} 0 \ . \tag{6.5}$$

Since $\hat{\theta}^*_{n,\text{aug}}$ is just an average of i.i.d. observations, it follows from the CLT that

$$\sqrt{n}\left(\hat{\theta}^*_{n,\text{aug}} - \theta\right) \xrightarrow{d} N\left(0, V^*\right) \ , \tag{6.6}$$

with V^* as defined in (6.2). Note that in order to characterize V^*, we used the fact that the three terms entering the expression for $\hat{\theta}^*_{n,\text{aug}}$ are uncorrelated. Whenever (6.5) holds, $\hat{\theta}_{n,\text{aug}}$ is also asymptotically normal with the same limiting variance V^* and, as a result, it is also a semiparametrically efficient estimator of θ.

6.5 Cross-fitting

In order to show that the AIPW estimator is semiparametrically efficient, we need (6.5) to hold. A convenient way to achieve this is through cross-fitting, which uses sample splitting to separate nuisance estimation from treatment-effect estimation. Roughly speaking, the nuisance functions are estimated on one part of the sample and then evaluated on a different part. In particular, it guarantees that the data we use to compute averages are independent of the data we use to estimate nuisance parameters. This separation reduces overfitting bias and makes the remainder terms easier to control.

In order to define cross-fitting, we introduce additional notation. First, split the data (at random) into two halves and denote the associated indices by $\mathcal{S}_1$ and $\mathcal{S}_2$. For the purposes of this class we focus on such simple partitions, but note that more generally one could consider K-fold partitions too. Next, denote by $\hat{\mu}^j_{n,a}(\cdot)$ and $\hat{\pi}^j_n(\cdot)$ the estimates of $\mu_a(\cdot)$ and $\pi(\cdot)$ obtained by using only the observations in the sample $\mathcal{S}_j$. Finally, let $\omega_j := |\mathcal{S}_j|/n$. Using this notation we can define the AIPW estimator via cross-fitting by

$$\hat{\theta}_{n,\text{aug}} = \omega_1 \hat{\theta}^1_n + \omega_2 \hat{\theta}^2_n \ , \tag{6.7}$$

where

$$\hat{\theta}^j_n \equiv \frac{1}{|\mathcal{S}_j|} \sum_{i \in \mathcal{S}_j} \left(\hat{\mu}^{-j}_{n,1}(W_i) - \hat{\mu}^{-j}_{n,0}(W_i) + A_i \frac{Y_i - \hat{\mu}^{-j}_{n,1}(W_i)}{\hat{\pi}^{-j}_n(W_i)} - (1 - A_i) \frac{Y_i - \hat{\mu}^{-j}_{n,0}(W_i)}{1 - \hat{\pi}^{-j}_n(W_i)} \right) .$$

Note that $\hat{\theta}^j_n$ is a treatment effect estimator on $\mathcal{S}_j$ that uses the data in $\mathcal{S}_{-j}$ to estimate its nuisance components.

This cross-estimation construction allows us to asymptotically ignore the idiosyncrasies of the specific estimators of $\mu_a(\cdot)$ and $\pi(\cdot)$ we chose to use, and to simply rely on the following high-level conditions:

1. **Overlap:** as defined in Assumption 6.1.

2. **Consistency:** All estimates are uniformly consistent,

$$\sup_{w \in \mathcal{W}} \left| \hat{\mu}^j_{n,a}(w) - \mu_a(w) \right| \xrightarrow{P} 0 \quad \text{and} \quad \sup_{w \in \mathcal{W}} \left| \hat{\pi}^j_n(w) - \pi(w) \right| \xrightarrow{P} 0 \ .$$

3. **Risk decay:** The product of the errors for the outcome and propensity models decays as

$$E\left[\left(\hat{\mu}^j_{n,a}(W) - \mu_a(W) \right)^2 \right] E\left[\left(\hat{\pi}^j_n(W) - \pi(W) \right)^2 \right] = o\left(\frac{1}{n} \right) \ .$$

Risk decay deserves further discussion. If both $\hat{\mu}_{n,a}$ and $\hat{\pi}_n$ attain the

parametric $\sqrt{n}$ rate, then the product of their mean-squared errors is of order $O(n^{-2})$, so the condition is automatically satisfied. Of course, this would typically require both models to be correctly specified.

More generally, the condition also holds if the nuisance estimators converge sufficiently fast in mean-squared error. For example, it is enough that the relevant estimators be $n^{1/4}$-consistent. This rate can be achieved by several ML methods under structured assumptions on $\mu_a(\cdot)$ and $\pi(\cdot)$, such as sparsity or other low-complexity conditions. Without such structure, however, the curse of dimensionality becomes binding and risk decay may fail.

Given these assumptions, we can show that (6.5) holds. To see this, first note that

$$\hat{\theta}^*_{n,\text{aug}} = \omega_1 \hat{\theta}^{1,*}_n + \omega_2 \hat{\theta}^{2,*}_n.$$

The star notation means that oracle nuisance components μ_a and π are used as in (6.4). Thus, $\hat{\theta}^{1,*}$ is analogous to (6.4) but only uses the half-sample $\mathcal{S}_1$.

Second, note that we can decompose $\hat{\theta}^1_n$ as

$$\begin{aligned}
\hat{\theta}^1_n &= \hat{m}^1_{n,1} - \hat{m}^1_{n,0}, \\
\hat{m}^1_{n,1} &= \frac{1}{|\mathcal{S}_1|} \sum_{i \in \mathcal{S}_1} \left(\hat{\mu}^2_{n,1}(W_i) + A_i \frac{Y_i - \hat{\mu}^2_{n,1}(W_i)}{\hat{\pi}^2_n(W_i)} \right) \\
\hat{m}^1_{n,0} &= \frac{1}{|\mathcal{S}_1|} \sum_{i \in \mathcal{S}_1} \left(\hat{\mu}^2_{n,0}(W_i) + (1 - A_i) \frac{Y_i - \hat{\mu}^2_{n,0}(W_i)}{1 - \hat{\pi}^2_n(W_i)} \right)
\end{aligned} \tag{6.8}$$

and define $\hat{m}^{1,*}_{n,0}$ and $\hat{m}^{1,*}_{n,1}$ analogously. Given this buildup, in order to verify (6.5), it suffices to show that

$$\sqrt{n} \left(\hat{m}^j_{n,a} - \hat{m}^{j,*}_{n,a} \right) \xrightarrow{P} 0 \ . \tag{6.9}$$

In order to do this, first decompose

$$\begin{aligned}
&\hat{m}^1_{n,1} - \hat{m}^{1,*}_{n,1} \\
&= \frac{1}{|\mathcal{S}_1|} \sum_{i \in \mathcal{S}_1} \left(\hat{\mu}^2_{n,1}(W_i) + A_i \frac{Y_i - \hat{\mu}^2_{n,1}(W_i)}{\hat{\pi}^2_n(W_i)} - \mu_1(W_i) - A_i \frac{Y_i - \mu_1(W_i)}{\pi(W_i)} \right) \\
&= \frac{1}{|\mathcal{S}_1|} \sum_{i \in \mathcal{S}_1} \left(\hat{\mu}^2_{n,1}(W_i) - \mu_1(W_i) \right) \left(1 - \frac{A_i}{\pi(W_i)} \right) \\
&\quad + \frac{1}{|\mathcal{S}_1|} \sum_{i \in \mathcal{S}_1} A_i \left(Y_i - \mu_1(W_i) \right) \left(\frac{1}{\hat{\pi}^2_n(W_i)} - \frac{1}{\pi(W_i)} \right) \\
&\quad - \frac{1}{|\mathcal{S}_1|} \sum_{i \in \mathcal{S}_1} A_i \left(\hat{\mu}^2_{n,1}(W_i) - \mu_1(W_i) \right) \left(\frac{1}{\hat{\pi}^2_n(W_i)} - \frac{1}{\pi(W_i)} \right) \ .
\end{aligned} \tag{6.10}$$

For the first term, conditional on $\mathcal{S}_2$, $\hat{\mu}^2_{(a)}$ is deterministic. Thus after conditioning on $\mathcal{S}_2$, the summands used to build this term become mean-zero

and independent. We prove convergence in L2, which implies convergence in probability by Chebyshev's inequality. Note that

$$\begin{aligned}
&E\left[\left(\frac{1}{|\mathcal{S}_1|}\sum_{i\in\mathcal{S}_1}\left(\hat{\mu}^2_{n,1}(W_i)-\mu_1(W_i)\right)\left(1-\frac{A_i}{\pi(W_i)}\right)\right)^2\right] \\
&= E\left[E\left[\left(\frac{1}{|\mathcal{S}_1|}\sum_{i\in\mathcal{S}_1}\left(\hat{\mu}^2_{n,1}(W_i)-\mu_1(W_i)\right)\left(1-\frac{A_i}{\pi(W_i)}\right)\right)^2 \mid \mathcal{S}_2\right]\right] \\
&= \frac{2}{n}E\left[E\left[\left(\hat{\mu}^2_{n,1}(W_i)-\mu_1(W_i)\right)^2\left(1-\frac{A_i}{\pi(W_i)}\right)^2 \mid \mathcal{S}_2\right]\right] \\
&= \frac{2}{n}E\left[E\left[\left(\hat{\mu}^2_{n,1}(W_i)-\mu_1(W_i)\right)^2\left(\frac{1}{\pi(W_i)}-1\right) \mid \mathcal{S}_2\right]\right] \\
&\leq \frac{2}{n\eta}E\left[\left(\hat{\mu}^2_{n,1}(W_i)-\mu_1(W_i)\right)^2\right] = \frac{o_P(1)}{n}\ ,
\end{aligned}$$

where the last step follows by consistency of $\hat{\mu}^2_{n,a}(\cdot)$ in mean-squared (which is implied by uniform consistency), $\mathcal{S}_1 \sim n/2$, and the overlap assumption. The key step in this argument was the second equality: because the summands become independent and mean-zero after conditioning, we "earn" a factor $|\mathcal{S}_1|^{-1}$, which is non-random and proportional to n^{-1}, due to concentration of i.i.d. sums. Note also that in the third equality we used the LIE conditional on W_i and worked out the square,

$$\left(1-\frac{A_i}{\pi(W_i)}\right)^2 = \left(1-2\frac{A_i}{\pi(W_i)}+\frac{A_i^2}{\pi(W_i)}\right)\ .$$

The second summand in our decomposition here can also be bounded similarly. Finally, for the last summand, use Cauchy-Schwarz:

$$\begin{aligned}
&\frac{1}{|\mathcal{S}_1|}\sum_{\{i:i\in\mathcal{S}_1,A_i=1\}}\left(\hat{\mu}^2_{n,1}(W_i)-\mu_1(W_i)\right)\left(\frac{1}{\hat{\pi}^2_n(W_i)}-\frac{1}{\pi(W_i)}\right) \\
&\quad\leq\sqrt{\frac{1}{|\mathcal{S}_1|}\sum_{\{i:i\in\mathcal{S}_1,A_i=1\}}\left(\hat{\mu}^2_{n,1}(W_i)-\mu_1(W_i)\right)^2} \\
&\qquad\times\sqrt{\frac{1}{|\mathcal{S}_1|}\sum_{\{i:i\in\mathcal{S}_1,A_i=1\}}\left(\frac{1}{\hat{\pi}^2_n(W_i)}-\frac{1}{\pi(W_i)}\right)^2} = o_P\left(\frac{1}{\sqrt{n}}\right)
\end{aligned}$$

by risk decay. Note that, due to consistency and overlap, the estimated propensities will eventually be uniformly bounded away from 0, $\eta/2 \leq \hat{\pi}^2_n(W_i) \leq 1-\eta/2$, and so the MSE for the inverse weights decays at the same rate as the MSE for the propensities themselves. This concludes the proof and shows that

$$\sqrt{n}\left(\hat{\theta}_{n,\text{aug}}-\hat{\theta}^*_{n,\text{aug}}\right)\xrightarrow{P}0\ .$$

6.6 Confidence Intervals

In general, this cross-fitting estimator can be done by splitting the data into K folds (above, K=2) and computing estimators $\hat{\mu}_{(a)}^{(-k)}(w)$, etc., excluding the k-th fold. Then, writing $k(i)$ as the mapping that takes an observation and puts it into one of the k folds, we can write

$$\hat{\theta}_{n,\text{aug}} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}_{n,1}^{(-k(i))}(W_i) - \hat{\mu}_{n,0}^{(-k(i))}(W_i) \right.$$
$$\left. + A_i\frac{Y_i - \hat{\mu}_{n,1}^{(-k(i))}(W_i)}{\hat{\pi}_n^{(-k(i))}(W_i)} - (1-A_i)\frac{Y_i - \hat{\mu}_{n,0}^{(-k(i))}(W_i)}{1-\hat{\pi}_n^{(-k(i))}(W_i)}\right) .$$

To do inference, note that the empirical variance of the efficient score converges to the efficient variance V^*:

$$\frac{1}{n-1}\sum_{i=1}^{n}\left(\mu_1(W_i) - \mu_0(W_i)\right.$$
$$\left. + A_i\frac{Y_i - \mu_1(W_i)}{\pi(W_i)} - (1-A_i)\frac{Y_i - \mu_0(W_i)}{1-\pi(W_i)} - \hat{\theta}^*_{n,\text{aug}}\right)^2 \xrightarrow{P} V^*$$

Our previous derivation then establishes that the same holds for cross-fitting: $\widehat{V}_{\text{aug}} \xrightarrow{P} V^*$, where

$$\widehat{V}_{\text{aug}} := \frac{1}{n-1}\sum_{i=1}^{n}\left(\hat{\mu}_{n,1}^{(-k(i))}(W_i) - \hat{\mu}_{n,0}^{(-k(i))}(W_i)\right.$$
$$\left. + A_i\frac{Y_i - \hat{\mu}_{n,1}^{(-k(i))}(W_i)}{\hat{\pi}_n^{(-k(i))}(W_i)} - (1-A_i)\frac{Y_i - \hat{\mu}_{n,0}^{(-k(i))}(W_i)}{1-\hat{\pi}_n^{(-k(i))}(W_i)} - \hat{\theta}_{n,\text{aug}}\right)^2 .$$

We can thus produce level-α confidence intervals for θ as

$$CS_n := \left[\hat{\theta}_{n,\text{aug}} \pm \frac{1}{\sqrt{n}}\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\sqrt{\widehat{V}_{\text{aug}}}\right]$$

where $\Phi(\cdot)$ is the standard Gaussian CDF.

6.7 Concluding Remarks

Today's lecture follows the notes by Stefan Wager, Wager [2020]. Additional resources on this topic can be found in Ding [2023] and the original papers, notably Robins et al. [1994].

6.8 Problems

Problem 6.1 *Prove* (6.6)

Problem 6.2 *Prove that the second term in* (6.10) *is* $o_P(n^{-1/2})$ *using similar steps as we showed for the first term.*

Problem 6.3 *Assume that* W *is discrete and takes* K *values. Show that* $\hat{\theta}_{n,\text{ipw}}$, $\hat{\theta}_{n,\text{pss}}$, $\hat{\theta}_{n,\text{mat}}$, $\hat{\theta}_{n,\text{reg}}$, *and* $\hat{\theta}_{n,\text{aug}}$, *where the unknown functions are estimated non-parametrically, are all numerically equivalent.*

Problem 6.4 *Revisiting Example 5.1, implement the regression estimator and the augmented IPW estimator. You may use the code in the last lecture for estimating the propensity score, and use a parametric (e.g., linear model) or non-parametric (e.g., tree) estimator of your choice for the outcome regressions. Compare your results with the estimates from the last lecture.*

Problem 6.5 *Show that the oracle AIPW estimator in* (6.4) *is unbiased for* θ. *In particular, prove that*

$$E[\hat{\theta}^*_{n,\text{aug}}] = \theta \ .$$

Problem 6.6 *Consider the population moment*

$$\psi(Y, A, W; \mu_1, \mu_0, \pi) := \mu_1(W) - \mu_0(W) + A\frac{Y - \mu_1(W)}{\pi(W)} - (1-A)\frac{Y - \mu_0(W)}{1 - \pi(W)} \ .$$

Show that under unconfoundedness and overlap,

$$E[\psi(Y, A, W; \mu_1, \mu_0, \pi)] = \theta \ .$$

Then show that this conclusion still holds if either:

(a) $\mu_1(W)$ *and* $\mu_0(W)$ *are replaced by arbitrary functions* $\bar{\mu}_1(W)$ *and* $\bar{\mu}_0(W)$, *while* $\pi(W)$ *is kept correctly specified; or*

(b) $\pi(W)$ *is replaced by an arbitrary function* $\bar{\pi}(W)$ *satisfying* $0 < \bar{\pi}(W) < 1$ *a.s., while* $\mu_1(W)$ *and* $\mu_0(W)$ *are kept correctly specified.*

Conclude that the population moment underlying AIPW is doubly robust.

Bibliography

P. Ding. A first course in causal inference. arXiv:2305.18793, 2023.

J. Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, pages 315–331, 1998.

K. Hirano, G. W. Imbens, and G. Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71 (4):1161–1189, 2003.

G. W. Imbens, W. K. Newey, and G. Ridder. Mean-square-error calculations for average treatment effects. 2006.

J. M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994.

S. Wager. Causal inference. Standford University, 2020.

Part II

Causality and Endogeneity

7

Endogeneity

Selection on observables is often difficult to justify in economic applications. Broadly speaking, unobserved confounders—such as preferences, private information, or expectations—tend to be the rule rather than the exception. In such settings, concerns about endogeneity, selection, and related issues naturally arise. More generally, if our goal is to identify the **causal effect** of a variable X on an outcome Y, the presence of unobserved confounders that are correlated with X (hence the name *confounders*) complicates matters. In these cases, it becomes difficult to disentangle whether the observed relationship is driven by X itself or by some omitted variable U, as illustrated in Figure 7.1 below.

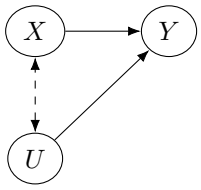


FIGURE 7.1: Unobserved factors are confounded with observed factors

Understanding the implications of endogeneity in our models is essential for being a good econometrician, whether theoretical or applied. We start our discussion of endogeneity within the context of the linear model, and then move on to more general settings.

7.1 Endogeneity in Linear Regression

Let (Y, X, U) be a random vector where Y and U take values in $\mathbf{R}$ and X takes values in $\mathbf{R}^{k+1}$. Assume further that the first component of X is constant and equal to one, i.e., $X = (X_0, X_1, \ldots, X_k)'$ with $X_0 = 1$. Let $\beta = (\beta_0, \beta_1, \ldots, \beta_k)' \in \mathbf{R}^{k+1}$ be such that

$$Y = X'\beta + U \ .$$

In contrast to our earlier discussion, we do not assume that $E[XU] = 0$. Any X_j such that $E[X_jU] = 0$ is said to be *exogenous*; any X_j such that $E[X_jU] \neq 0$ is said to be *endogenous*. By normalizing β_0 if necessary, we assume X_0 is exogenous. We interpret this regression as a causal model.

Note that since $E[XU] \neq 0$ we have that

$$E[XY] = E[XX']\beta + E[XU]$$

and so

$$E[XX']^{-1}E[XY] = \beta + E[XX']^{-1}E[XU] \ .$$

The results from the previous class showed that the least squares estimator $\hat{\beta}_n$ of β converges to $E[XX']^{-1}E[XY]$. It follows that

$$\hat{\beta}_n \xrightarrow{P} \beta + E[XX']^{-1}E[XU] \ , \tag{7.1}$$

and is therefore inconsistent for β under endogeneity.

We now briefly review some common ways in which endogeneity may arise. In many of these cases, we focus on the simple case with a single regressor ($k = 1$), where the model takes the form

$$Y = \beta_0 + \beta_1 X_1 + U \ ,$$

and the asymptotic bias in (7.1) simplifies to

$$\hat{\beta}_{1,n} \xrightarrow{P} \beta_1 + \frac{\text{Cov}[X_1, U]}{\text{Var}[X_1]} \ . \tag{7.2}$$

7.1.1 Omitted Variables

Suppose $k = 2$, so

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U \ .$$

We are interpreting this regression as a causal model and are willing to assume that $E[XU] = 0$ (i.e., $E[U] = E[X_1U] = E[X_2U] = 0$), but X_2 is unobserved. An example of a situation like this is when Y is wages, X_1 is education, and X_2 is ability. Given unobserved ability, we may rewrite this model as

$$Y = \beta_0^* + \beta_1^* X_1 + U^* \ ,$$

with

$$\begin{aligned} \beta_0^* &= \beta_0 + \beta_2 E[X_2] \\ \beta_1^* &= \beta_1 \\ U^* &= \beta_2(X_2 - E[X_2]) + U \ . \end{aligned}$$

Note that we have normalized β_0^* so that $E[U^*] = 0$. In this model,

$$E[X_1U^*] = \beta_2 \text{Cov}[X_1, X_2] \ ,$$

so X_1 is endogenous whenever $\beta_2 \neq 0$ and $\mathrm{Cov}[X_1, X_2] \neq 0$. Based on the results from the previous class, it follows immediately that running a regression of Y on X_1 produces an estimator with the property that

$$\hat{\beta}^*_{1,n} \xrightarrow{P} \beta_1 + \beta_2 \frac{\mathrm{Cov}[X_1, X_2]}{\mathrm{Var}[X_1]} \ , \tag{7.3}$$

where the term $\beta_2[\mathrm{Var}[X_1]]^{-1}\mathrm{Cov}[X_1, X_2]$ is usually referred to as *omitted variable bias*.

7.1.2 Measurement Error

Partition X into X_0 and X_1, where $X_0 = 1$ and X_1 takes values in $\mathbf{R}^k$. Partition β analogously. In this notation,

$$Y = \beta_0 + X_1'\beta_1 + U \ .$$

We are interpreting this regression as a causal model and are willing to assume that $E[XU] = 0$, but X_1 is *not* observed. Instead, $\hat{X}_1$ is observed, where

$$\hat{X}_1 = X_1 + V \ .$$

Assume $E[V] = 0$, $\mathrm{Cov}[X_1, V] = 0$, and $\mathrm{Cov}[U, V] = 0$. We may therefore rewrite this model as

$$Y = \beta^*_0 + \hat{X}_1'\beta^*_1 + U^* \ ,$$

with

$$\begin{aligned} \beta^*_0 &= \beta_0 \\ \beta^*_1 &= \beta_1 \\ U^* &= -V'\beta_1 + U \ . \end{aligned}$$

In this model,

$$E[\hat{X}_1 U^*] = -E[\hat{X}_1 V'\beta_1] = -E[VV']\beta_1 \ ,$$

so $\hat{X}_1$ is typically endogenous. Note that in the case where X_1 is a scalar random variable, and using results from the previous class, it follows that running a regression of Y on $\hat{X}_1$ produces an estimator with the property that

$$\hat{\beta}^*_{1,n} \xrightarrow{P} \beta_1 + \frac{E[\hat{X}_1 U^*]}{\mathrm{Var}[\hat{X}_1]} = \beta_1 \left(1 - \frac{\mathrm{Var}[V]}{\mathrm{Var}[\hat{X}_1]}\right) < \beta_1 \ , \tag{7.4}$$

so that the regression coefficient is biased towards zero when the regressor of interest is measured with the so-called classical random errors. The last inequality follows from using $\hat{X}_1 = X_1 + V$. Indeed, in the extreme case where $\hat{X}_1 = V$, it follows that $\hat{\beta}^*_{1,n} \xrightarrow{P} 0$.

7.1.3 Simultaneity

A classical source of endogeneity arises when the variable of interest is itself jointly determined with the outcome. The canonical example is the determination of prices and quantities in a market. In competitive markets, price is not externally assigned and then taken as fixed by buyers and sellers. Instead, price and quantity are equilibrium outcomes determined by the interaction of supply and demand. This is one of the classical motivations for simultaneous-equation models in econometrics.

Historically, this issue played a central role in the development of instrumental variables. Early empirical work attempted to learn about demand from observed correlations between prices and quantities, but it soon became clear that such correlations need not trace out a demand curve. In particular, once both supply and demand are allowed to shift, observed market outcomes reflect the intersection of two behavioral relationships rather than movement along a single one. This identification problem was emphasized in the early simultaneous-equations literature and later addressed by introducing variables that shift one side of the market but not the other. Work along these lines dates back at least to Engel's (1861) study of demand, Hooker's (1905) early correlation of prices and quantities, and Lenoir's (1913) distinction between statistical price–quantity correlations and a structural demand curve. Tinbergen (1930) is often credited with one of the earliest solutions to the identification problem in supply-and-demand systems using excluded variables.

To see the problem formally, denote by Q^d the quantity demanded and by Q^s the quantity supplied. As functions of a hypothetical non-market-clearing price $\tilde{P}$, suppose that

$$\begin{aligned} Q^d &= \beta_0^d + \beta_1^d \tilde{P} + U^d \\ Q^s &= \beta_0^s + \beta_1^s \tilde{P} + U^s \ , \end{aligned}$$

where $E[U^d] = E[U^s] = E[U^d U^s] = 0$. We think of U^d and U^s as unobserved demand and supply shocks, respectively. For example, U^d may capture changes in tastes or income, while U^s may capture changes in weather, productivity, or input costs.

We observe (Q, P), where Q and P are such that the market clears, i.e.,

$$Q^d = Q^s \ .$$

Therefore,

$$\beta_0^d + \beta_1^d P + U^d = \beta_0^s + \beta_1^s P + U^s \ ,$$

which implies

$$P = \frac{1}{\beta_1^s - \beta_1^d} \left(\beta_0^d - \beta_0^s + U^d - U^s \right) \ .$$

This expression makes the source of endogeneity transparent. A positive demand shock U^d raises the equilibrium price, so price is positively correlated with the unobserved disturbance in the demand equation. Likewise, a positive

supply shock U^s lowers the equilibrium price, so price is negatively correlated with the unobserved disturbance in the supply equation. Thus, price is endogenous in both equations

$$\begin{aligned} Q &= \beta_0^d + \beta_1^d P + U^d \\ Q &= \beta_0^s + \beta_1^s P + U^s \end{aligned}$$

because

$$\begin{aligned} E[PU^d] &= \frac{\text{Var}[U^d]}{\beta_1^s - \beta_1^d} \\ E[PU^s] &= -\frac{\text{Var}[U^s]}{\beta_1^s - \beta_1^d} \ . \end{aligned}$$

As in the previous examples, OLS therefore fails. A regression of Q on P will not recover either β_1^d or β_1^s. The reason is that the observed pairs (P, Q) combine movements along the demand curve with shifts in the demand curve, and likewise combine movements along the supply curve with shifts in the supply curve. In other words, equilibrium data alone do not generally allow us to separate demand from supply.

This insight points directly toward instrumental variables. To identify the demand curve, for example, one would like to observe variables that shift supply but do not directly affect demand. Such variables generate variation in price that is relevant for quantity demanded but exogenous to the demand disturbance. Similarly, to identify the supply curve, one would like to observe variables that shift demand but do not directly affect supply. We return to this logic in the next section.

7.2 Instrumental Variables

To address the challenge posed by endogeneity, that is, the failure of the condition $E[XU] = 0$ in the model

$$Y = X'\beta + U \ ,$$

we introduce an additional random vector called *instruments*, denoted by Z. Specifically, we assume:

$$Z = (Z_0, Z_1, \ldots, Z_\ell)' \in \mathbf{R}^{\ell+1}, \quad \text{with } \ell + 1 \geq k + 1 \ .$$

We assume that any exogenous component of X is contained in Z (the so-called included instruments). In particular, we assume that the first component of Z is constant and equal to one, i.e., $Z = (Z_0, Z_1, \ldots, Z_\ell)'$ with $Z_0 = 1$. However, since there are endogenous components in X, it must be the case

that Z includes variables that are *not* components of X (the so-called excluded instruments).

We impose the following conditions on Z:

Instrument Exogeneity: The instruments are uncorrelated with the structural error term U:

$$E[ZU] = 0 \ .$$

This ensures that any correlation between Z and Y operates through X, not through omitted variables in U.

Instrument Relevance (Rank Condition): The instruments must be sufficiently correlated with the endogenous regressors. Specifically, we require:

$$\text{rank}(E[ZX']) = k + 1 \ .$$

This ensures that the system of moment conditions has a unique solution for β.

Order Condition: A necessary condition for relevance is that the number of instruments (excluding the intercept) is at least as large as the number of endogenous regressors:

$$\ell \geq k \ .$$

Regularity Conditions: We also assume the following expectations are finite:

$$E[ZX'] < \infty \quad \text{and} \quad E[ZZ'] < \infty \ ,$$

and that there is no perfect collinearity among the components of Z.

These properties are essential for identification of the (homogeneous) causal effect β in the presence of endogeneity.

Using the fact that $U = Y - X'\beta$ and $E[ZU] = 0$, we see that β solves the system of equations

$$E[ZY] = E[ZX']\beta \ .$$

Since $\ell + 1 \geq k + 1$, this may be an over-determined system of equations. In order to find an explicit formula for β, the following lemma is useful.

Lemma 7.1 *Suppose there is no perfect collinearity in Z and let Π be such that $BLP(X|Z) = \Pi'Z$. $E[ZX']$ has rank $k+1$ if and only if Π has rank $k+1$. Moreover, the matrix $\Pi'E[ZX']$ is invertible.*

PROOF: Write $X = \Pi'Z + V$ where $E[ZV'] = 0$. It follows that $E[ZX'] = E[ZZ']\Pi$. Recall the rank inequality, which states that

$$\text{rank}(AB) \leq \min\{\text{rank}(A), \text{rank}(B)\}$$

for any conformable matrices A and B. Applying this result, we see that

$$\operatorname{rank}(E[ZZ']\Pi) \leq \operatorname{rank}(\Pi) \ .$$

We have further that

$$\operatorname{rank}(\Pi) = \operatorname{rank}(E[ZZ']^{-1}E[ZZ']\Pi) \leq \operatorname{rank}(E[ZZ']\Pi) \ .$$

Hence,

$$\operatorname{rank}(E[ZX']) = \operatorname{rank}(E[ZZ']\Pi) = \operatorname{rank}(\Pi) \ ,$$

as desired.

To complete the proof, note that $\Pi' E[ZX'] = \Pi' E[ZZ']\Pi$ and argue that $\Pi' E[ZZ']\Pi$ is invertible using arguments given earlier. ■

Since β solves $\Pi' E[ZY] = \Pi' E[ZX']\beta$, we arrive at two formulae for β by applying the lemma:

$$\beta = (\Pi' E[ZX'])^{-1}\Pi' E[ZY] \tag{7.5}$$

$$= (\Pi' E[ZZ']\Pi)^{-1}\Pi' E[ZY] \ . \tag{7.6}$$

Note that if $k = \ell$, then Π is an invertible matrix and therefore

$$\beta = (E[ZX'])^{-1}E[ZY] \ . \tag{7.7}$$

In this case, we say that β is *exactly identified*. Otherwise, we say that β is *over-identified*.

A third formula for β arises by replacing Π with $E[ZZ']^{-1}E[ZX']$,

$$\beta = \left(E[ZX']' E[ZZ']^{-1} E[ZX']\right)^{-1} E[ZX']' E[ZZ']^{-1} E[ZY] \ . \tag{7.8}$$

Before proceeding, it is useful to use the preceding lemma to further examine the rank condition in some simpler settings. To this end, consider the case where $k = \ell$ and only X_k is endogenous. Let $Z_j = X_j$ for all $0 \leq j \leq k-1$. In this case,

$$\Pi' = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \\ \pi_0 & \pi_1 & \cdots & \pi_{\ell-1} & \pi_\ell \end{pmatrix}$$

The rank condition therefore requires $\pi_\ell \neq 0$. In other words, the instrument Z_ℓ must be "correlated with X_k after controlling for $X_0, X_1, \ldots, X_{k-1}$."

7.2.1 Partition of β: solve for endogenous components

Partition X into X_1 and X_2, where X_2 is exogenous. Partition Z into Z_1 and Z_2 and β into β_1 and β_2 analogously. Note that $Z_2 = X_2$ are *included* instruments and Z_1 are *excluded* instruments. In this notation,

$$Y = X_1'\beta_1 + X_2'\beta_2 + U \ .$$

Note that

$$\begin{aligned} \text{BLP}(Y|Z_2) &= \text{BLP}(X_1'\beta_1|Z_2) + \text{BLP}(X_2'\beta_2|Z_2) + \text{BLP}(U|Z_2) \\ &= \text{BLP}(X_1|Z_2)'\beta_1 + X_2'\beta_2 \ , \end{aligned}$$

where the second equality uses the fact that $E[Z_2U] = 0$. It follows that

$$Y^* = X_1^{*\prime}\beta_1 + U \ ,$$

where

$$\begin{aligned} Y^* &= Y - \text{BLP}(Y|Z_2) \\ X_1^* &= X_1 - \text{BLP}(X_1|Z_2) \ . \end{aligned}$$

This calculation shows again the sense in which we may interpret β_1 as summarizing the effect of X_1 on Y "after controlling for X_2." In the exactly identified case, it follows that

$$E[Z_1Y^*] = E[Z_1X_1^{*\prime}]\beta_1 \ .$$

Since there must be a unique solution to this system of equations, it must be the case that $E[Z_1X_1^{*\prime}]$ is invertible. It follows that

$$\beta_1 = E[Z_1X_1^{*\prime}]^{-1}E[Z_1Y^*] \ .$$

In the over-identified case, we may repeat this calculation with $\hat{X}_1^* = \text{BLP}(X_1^*|Z_1)$ in place of Z_1. This yields

$$\begin{aligned} \beta_1 &= E[\hat{X}_1^*X_1^{*\prime}]^{-1}E[\hat{X}_1^*Y^*] \\ &= E[\hat{X}_1^*\hat{X}_1^{*\prime}]^{-1}E[\hat{X}_1^*Y^*] \ , \end{aligned}$$

where the second equality uses the fact that $X_1^* = \hat{X}_1^* + V$ with $E[\hat{X}_1^*V'] = 0$.

7.3 Estimating β

Let (Y, X, Z, U) be a random vector where Y and U take values in $\mathbf{R}$, X takes values in $\mathbf{R}^{k+1}$, and Z takes values in $\mathbf{R}^{\ell+1}$. Let $\beta \in \mathbf{R}^{k+1}$ be such that

$$Y = X'\beta + U \ .$$

Suppose $E[ZX'] < \infty$, $E[ZZ'] < \infty$, $E[ZU] = 0$, there is no perfect collinearity in Z, and that the rank of $E[ZX']$ is $k+1$. We now discuss estimation of β.

7.3.1 The Instrumental Variables (IV) Estimator

We first consider the case in which $k = \ell$. Let (Y, X, Z, U) be distributed as described above and denote by P the marginal distribution of (Y, X, Z). Let $(Y_1, X_1, Z_1), \ldots, (Y_n, X_n, Z_n)$ be an i.i.d. sequence of random variables with distribution P. By analogy with the expression we derived for β in (7.7) under these assumptions, the natural estimator of β is simply

$$\hat{\beta}_n = \left(\frac{1}{n} \sum_{1 \le i \le n} Z_i X_i' \right)^{-1} \left(\frac{1}{n} \sum_{1 \le i \le n} Z_i Y_i \right) .$$

This estimator is called the *instrumental variables* (IV) estimator of β. Note that $\hat{\beta}_n$ satisfies

$$\frac{1}{n} \sum_{1 \le i \le n} Z_i (Y_i - X_i' \hat{\beta}_n) = 0 .$$

In particular, $\hat{U}_i = Y_i - X_i' \hat{\beta}_n$ satisfies

$$\frac{1}{n} \sum_{1 \le i \le n} Z_i \hat{U}_i = 0 .$$

To gain further insight into the IV estimator, partition X into X_0 and X_1, where $X_0 = 1$ and X_1 is assumed to take values in $\mathbf{R}$. Do the same with Z and β. An interesting interpretation of the IV estimator of β_1 is obtained by multiplying and dividing by $\frac{1}{n} \sum_{i=1}^n (Z_{1,i} - \bar{Z}_{1,n})^2$, i.e.,

$$\hat{\beta}_{1,n} = \frac{\frac{1}{n} \sum_{i=1}^n (Z_{1,i} - \bar{Z}_{1,n}) Y_i / \frac{1}{n} \sum_{i=1}^n (Z_{1,i} - \bar{Z}_{1,n})^2}{\frac{1}{n} \sum_{i=1}^n (Z_{1,i} - \bar{Z}_{1,n}) X_{1,i} / \frac{1}{n} \sum_{i=1}^n (Z_{1,i} - \bar{Z}_{1,n})^2} . \tag{7.9}$$

The IV estimator of β_1 is simply the ratio of the regression slope of Y on Z_1 (the so-called reduced form) to the regression slope of X_1 on Z_1 (the so-called first stage). To see this in a different way, write the model as

$$Y = \beta_0 + \beta_1 X_1 + U$$

and

$$X_1 = \pi_0 + \pi_1 Z_1 + V ,$$

so that substituting the second equation into the first delivers

$$Y = \beta_0^* + \beta_1 \pi_1 Z_1 + U^*$$

with

$$\begin{aligned} \beta_0^* &= \beta_0 + \beta_1 \pi_0 \\ U^* &= U + \beta_1 V . \end{aligned}$$

Thus, the estimated slope in the reduced form converges in probability to $\beta_1\pi_1$, while the estimated slope in the first stage converges to π_1. The IV estimator takes the ratio of these two, therefore delivering a consistent estimator of β_1. Note that the IV estimand is predicated on the notion that the first stage slope is not zero ($\pi_1 \neq 0$), which is just another way to state our rank condition in this simple case.

This estimator may be expressed more compactly using matrix notation. Define

$$\begin{aligned} \mathbb{Z} &= (Z_1, \ldots, Z_n)' \\ \mathbb{X} &= (X_1, \ldots, X_n)' \\ \mathbb{Y} &= (Y_1, \ldots, Y_n)' \ . \end{aligned}$$

In this notation, we have

$$\hat{\beta}_n = (\mathbb{Z}'\mathbb{X})^{-1}(\mathbb{Z}'\mathbb{Y}) \ .$$

7.3.2 The Two-Stage Least Squares (TSLS) Estimator

Now consider the case in which $\ell > k$. The expressions derived for β in this case involved Π, where $\text{BLP}(X|Z) = \Pi'Z$. An estimate of Π can be obtained by OLS. More precisely, since $\Pi = E[ZZ']^{-1}E[ZX']$, a natural estimator of Π is

$$\hat{\Pi}_n = \left(\frac{1}{n} \sum_{1 \leq i \leq n} Z_i Z_i' \right)^{-1} \left(\frac{1}{n} \sum_{1 \leq i \leq n} Z_i X_i' \right) \ .$$

With this estimator of Π, a natural estimator of β is simply

$$\begin{aligned} \hat{\beta}_n &= \left(\frac{1}{n} \sum_{1 \leq i \leq n} \hat{\Pi}_n' Z_i X_i' \right)^{-1} \left(\frac{1}{n} \sum_{1 \leq i \leq n} \hat{\Pi}_n' Z_i Y_i \right) \\ &= \left(\frac{1}{n} \sum_{1 \leq i \leq n} \hat{\Pi}_n' Z_i Z_i' \hat{\Pi}_n \right)^{-1} \left(\frac{1}{n} \sum_{1 \leq i \leq n} \hat{\Pi}_n' Z_i Y_i \right) \ . \end{aligned}$$

The first equation above provides an interpretation of the TSLS estimator as an IV estimator with $\hat{\Pi}_n' Z_i$ playing the role of the instrument. Note further that if $k + 1 = \ell + 1$ and $\hat{\Pi}_n$ is invertible, then the TSLS estimator of β is exactly equal to the IV estimator of β. The second equality might be expected from our calculations in (7.6). To justify it here, write $X_i = \hat{\Pi}_n' Z_i + \hat{V}_i$ and note from properties of OLS that

$$\frac{1}{n} \sum_{1 \leq i \leq n} Z_i \hat{V}_i' = 0 \ .$$

This estimator of β is called the *two-stage least squares* (TSLS) estimator of β. Note that $\hat{\beta}_n$ satisfies

$$\frac{1}{n}\sum_{1\leq i\leq n}\hat{\Pi}_n' Z_i(Y_i - X_i'\hat{\beta}_n) = 0 \ .$$

In particular, $\hat{U}_i = Y_i - X_i'\hat{\beta}_n$ satisfies

$$\frac{1}{n}\sum_{1\leq i\leq n}\hat{\Pi}_n' Z_i\hat{U}_i = 0 \ .$$

Notice that this implies that $\hat{U}_i$ is orthogonal to all instruments that coincide with the exogenous regressors, but may not be orthogonal to the other regressors. It is termed the TSLS estimator because it may be obtained in the following way: first, regress (each component of) X_i on Z_i to obtain $\hat{X}_i = \hat{\Pi}_n' Z_i$; second, regress Y_i on $\hat{X}_i$ to obtain $\hat{\beta}_n$. However, in order to obtain proper standard errors, it is recommended to compute the estimator in one step (see the following lecture).

The estimator may again be expressed more compactly using matrix notation. Define

$$\begin{aligned}\hat{\mathbb{X}} &= (\hat{X}_1, \ldots, \hat{X}_n)' \\ &= \mathbb{P}_Z\mathbb{X} \ ,\end{aligned}$$

where

$$\mathbb{P}_Z = \mathbb{Z}(\mathbb{Z}'\mathbb{Z})^{-1}\mathbb{Z}'$$

is the projection matrix onto the column space of $\mathbb{Z}$. In this notation, we have

$$\begin{aligned}\hat{\beta}_n &= (\hat{\mathbb{X}}'\mathbb{X})^{-1}(\hat{\mathbb{X}}'\mathbb{Y}) \\ &= (\mathbb{X}'\mathbb{P}_Z\mathbb{X})^{-1}(\mathbb{X}'\mathbb{P}_Z\mathbb{Y}) \ ,\end{aligned}$$

which should be expected given our previous derivation in (7.8).

7.4 Some examples of IVs in practice

Instrumental variables are widely used in applied work to address endogeneity concerns, and over time a variety of instruments have been proposed across different fields. Table 7.1 collects a few well-known examples. Here we briefly discuss three of them to illustrate how the key IV conditions—relevance and exogeneity—are evaluated in context.

In their study on the effect of fertility on female labor supply, Angrist and Evans [1998] use the occurrence of twins at the second birth as an instrument (though this is not the instrument they recommend in the end). The

treatment A (the endogenous X_1 in our notation) is an indicator for having more than two children, and the outcome Y is a labor market outcome for the mother. The idea is that having twins increases fertility in a way that is plausibly exogenous to labor supply decisions. While the instrument appears relevant—twins mechanically increase the likelihood of having more than two children—its exogeneity has been questioned. For instance, twinning rates vary with maternal age, and twins affect not only the number but also the spacing of children, which may have independent effects on labor outcomes.

A cleaner example comes from Angrist [1990], who studies the causal effect of military service on later-life outcomes using the Vietnam draft lottery as an instrument. The treatment A is veteran status, and the instrument Z is a dummy for receiving a low lottery number. The instrument is relevant: individuals with low numbers were more likely to serve. In addition, since lottery numbers were randomly assigned, it is tempting to conclude that Z is exogenous. However, exogeneity is a stronger condition than randomness: it requires that Z be independent of unobserved determinants of the outcome. In this case, there would be concerns if individuals reacted to their draft status — by enrolling in school, moving abroad, or otherwise altering life decisions—which could create correlation between the instrument and U.

A third example is from Sarsons [2015], who studies the relationship between income shocks and religious conflict in India. Rainfall shocks are used as an instrument for income, leveraging the fact that agricultural income is sensitive to rain. Exogeneity is often assumed on the basis that rainfall is naturally random, but Sarsons shows that this assumption may not hold. In particular, she finds that even in districts with irrigation dams—where income is largely insulated from rainfall—rainfall still predicts conflict. This suggests that rainfall may affect conflict through non-income channels, violating the exclusion restriction (or the exogeneity condition).

These examples highlight the central challenge of IV analysis: while relevance is often straightforward to test, exogeneity and exclusion must be argued carefully based on institutional knowledge and empirical evidence.

7.5 Concluding Remarks

The material in this chapter borrows from notes by Azeem Shaikh that he kindly shared with me. In general, the topic of linear IV and endogeneity is covered in most standard sources, including Hansen [2022] and Wooldridge [2010].

TABLE 7.1: Examples of instruments used in practice

Outcome Variable	Endogenous Variable	Instrumental Variable(s)	Reference
Labor supply	Disability insurance replacement rates	Region and time variation in benefit rules	Gruber (2000)
Labor supply	Fertility	Sibling sex composition	Angrist and Evans (1998)
Education, Labor supply	Out-of-wedlock fertility	Occurrence of twin births	Bronars and Grogger (1994)
Wages	Unemployment insurance tax rate	State laws	Anderson and Meyer (2000)
Earnings	Years of schooling	Region and time variation in school construction	Duflo (2001)
Earnings	Years of schooling	Proximity to college	Card (1995)
Earnings	Years of schooling	Quarter of birth	Angrist and Krueger (1991)
Earnings	Veteran status	Cohort dummies	Imbens and van der Klaauw (1995)
Earnings	Veteran status	Draft lottery number	Angrist (1990)
Achievement test scores	Class size	Discontinuities in class size due to maximum class-size rule	Angrist and Lavy (1999)
College enrollment	Financial aid	Discontinuities in financial aid formula	van der Klaauw (1996)
Health	Heart attack surgery	Proximity to cardiac care centers	McClellan, McNeil and Newhouse (1994)
Crime	Police	Electoral cycles	Levitt (1997)
Employment and Earnings	Length of prison sentence	Randomly assigned federal judges	Kling (1999)
Birth weight	Maternal smoking	State cigarette taxes	Evans and Ringel (1999)

7.6 Problems

Problem 7.1 *Let X and Z be $k+1$-dimensional random vectors. Suppose the rank of $E[ZX']$ is $k+1$. Show that there is no perfect collinearity in Z.*

Problem 7.2 *Consider the measurement error setting in Section 7.1.2. Suppose $\beta_1 > 0$. Show that $\frac{\text{Cov}[\hat{X},Y]}{\text{Var}[\hat{X}]} \leq \beta_1 \leq \frac{\text{Var}[Y]}{\text{Cov}[\hat{X},Y]}$. Interpret the upper and lower bounds in terms of coefficients from a regression.*

Problem 7.3 *Consider the setting in Section 7.2, let $V = X - \text{BLP}(X \mid Z)$ and define $V'\gamma = \text{BLP}(U \mid V)$. Rewrite the original model $Y = X'\beta + U$ as*

$$Y = X'\beta + V'\gamma + \tilde{U},$$

where $\tilde{U} = U - V'\gamma$. Show that in this model, X and V are exogenous. (The variable V is called a control variable.) Use the results on solving for subvectors in linear regression to derive an expression for β. Show that this expression is equal to the one derived in class.

Problem 7.4 *Consider the setting in Section 7.2.1. Show that*

$$\beta_1 = E[Z_1^* X_1{}']^{-1} E[Z_1^* Y] \ ,$$

where $Z_1^* = Z_1 - BLP(Z_1 \mid Z_2)$.

Problem 7.5 *Angrist and Evans (1998)Angrist and Evans [1998] studied the relationship between fertility decisions and female labor supply. Read the paper, explain why fertility decisions are considered endogenous, what instruments are used, and how the validity of the instruments is justified.*

Problem 7.6 *Consider the exactly identified IV model with one endogenous regressor and one excluded instrument:*

$$Y = \beta_0 + \beta_1 X + U \ ,$$

where the instrument is Z. *Suppose* $E[ZU] = 0$ *and* $\mathrm{Cov}(Z, X) \neq 0$.

(a) Show that

$$\beta_1 = \frac{\mathrm{Cov}(Z, Y)}{\mathrm{Cov}(Z, X)} \ .$$

(b) Show that if $\mathrm{Cov}(Z, X) = 0$, *then* β_1 *is not identified from the moment condition* $E[Z(Y - \beta_0 - \beta_1 X)] = 0$.

(c) Interpret the formula in part (a) in terms of the reduced form and first stage.

Problem 7.7 *Consider the supply and demand model in Section 7.1.3:*

$$\begin{aligned} Q^d &= \beta_0^d + \beta_1^d \tilde{P} + U^d \\ Q^s &= \beta_0^s + \beta_1^s \tilde{P} + U^s \ . \end{aligned}$$

Suppose now that there is an observed variable Z *that shifts supply but not demand, so that the model becomes*

$$\begin{aligned} Q^d &= \beta_0^d + \beta_1^d \tilde{P} + U^d \\ Q^s &= \beta_0^s + \beta_1^s \tilde{P} + \gamma Z + U^s \ , \end{aligned}$$

where $E[ZU^d] = E[ZU^s] = 0$ *and* $\gamma \neq 0$.

(a) Solve for the equilibrium price P *as a function of* Z, U^d, *and* U^s.

(b) Show that Z *is correlated with* P.

(c) Explain why Z *is a valid instrument for identifying the demand equation.*

(d) What economic interpretation would you give to such a variable Z*?*

Problem 7.8 *Consider the linear model*

$$Y = \beta_0 + \beta_1 X + U \ ,$$

where X is endogenous. Suppose there exists a random variable V such that

$$E[U \mid X, V] = E[U \mid V] \ .$$

Define

$$g(V) := E[U \mid V] \ ,$$

and rewrite the model as

$$Y = \beta_0 + \beta_1 X + g(V) + \tilde{U} \ ,$$

where

$$\tilde{U} := U - g(V) \ .$$

(a) Show that

$$E[\tilde{U} \mid X, V] = 0 \ .$$

(b) Explain why, after controlling for V, the regressor X is exogenous in the regression above.

(c) Suppose now that

$$g(V) = \gamma V$$

for some scalar γ. Show that β_1 can be identified as the coefficient on X in the linear regression of Y on $(1, X, V)$.

(d) Briefly relate this result to the omitted-variables problem discussed earlier in the chapter.

Bibliography

J. D. Angrist. Lifetime earnings and the vietnam era draft lottery: evidence from social security administrative records. *The american economic review*, pages 313–336, 1990.

J. D. Angrist and W. N. Evans. Children and their parents' labor supply: Evidence from exogenous variation in family size. *American Economic Review*, 88(3):450–477, 1998.

B. Hansen. *Econometrics*. Princeton University Press, 2022.

H. Sarsons. Rainfall and conflict: A cautionary tale. *Journal of Development Economics*, 115:62–72, 2015. doi: 10.1016/j.jdeveco.2014.12.007.

J. M. Wooldridge. *Econometric analysis of cross section and panel data*. MIT press, 2010.

8

Properties of Two Stages Least Squares

In the previous lecture, we introduced the TSLS estimand and its sample analogue. In this chapter, we study the large-sample properties of the resulting estimator, discuss its efficiency relative to other IV estimators, and examine two important special cases: binary endogenous variables and weak instruments.

8.1 Properties of the TSLS Estimator

Let (Y, X, Z, U) be a random vector where Y and U take values in $\mathbf{R}$, X takes values in $\mathbf{R}^{k+1}$, and Z takes values in $\mathbf{R}^{\ell+1}$. Let $\beta \in \mathbf{R}^{k+1}$ be such that

$$Y = X'\beta + U \ .$$

Suppose that $E[ZX'] < \infty$, $E[ZZ'] < \infty$, and $E[ZU] = 0$. Assume further that there is no perfect collinearity in Z and that the rank of $E[ZX']$ is $k+1$. Denote by P the marginal distribution of (Y, X, Z). Let $(Y_1, X_1, Z_1), \ldots, (Y_n, X_n, Z_n)$ be an i.i.d. sequence of random variables with distribution P. Above we described estimation of β via TSLS under these assumptions. We now discuss properties of the resulting estimator, $\hat{\beta}_n$, imposing stronger assumptions as needed.

8.1.1 Consistency

Under the assumptions stated above, the TSLS estimator, $\hat{\beta}_n$, is consistent for β, i.e., $\hat{\beta}_n \xrightarrow{P} \beta$ as $n \to \infty$. To see this, first recall from our results on OLS that

$$\hat{\Pi}_n \xrightarrow{P} \Pi$$

as $n \to \infty$. Next, note that the WLLN implies that

$$\frac{1}{n} \sum_{1 \le i \le n} Z_i Z_i' \xrightarrow{P} E[ZZ']$$

$$\frac{1}{n} \sum_{1 \le i \le n} Z_i Y_i \xrightarrow{P} E[ZY]$$

as $n \to \infty$. The desired result therefore follows from the CMT.

8.1.2 Limiting Distribution

In addition to the assumptions above, assume that $\text{Var}[ZU] = E[ZZ'U^2] < \infty$. Then,

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} N(0, \mathbb{V})$$

as $n \to \infty$, where

$$\mathbb{V} = E[\Pi' Z Z' \Pi]^{-1} \Pi' \text{Var}[ZU] \Pi E[\Pi' Z Z' \Pi]^{-1} \ .$$

To see this, note that

$$\sqrt{n}(\hat{\beta}_n - \beta) = \left(\hat{\Pi}_n' \left(\frac{1}{n} \sum_{1 \le i \le n} Z_i Z_i' \right) \hat{\Pi}_n \right)^{-1} \left(\hat{\Pi}_n' \left(\frac{1}{\sqrt{n}} \sum_{1 \le i \le n} Z_i U_i \right) \right) \ .$$

As in the preceding section, we have that

$$\begin{aligned} \hat{\Pi}_n &\xrightarrow{P} \Pi \\ \frac{1}{n} \sum_{1 \le i \le n} Z_i Z_i' &\xrightarrow{P} E[ZZ'] \end{aligned}$$

as $n \to \infty$. The CLT implies that

$$\frac{1}{\sqrt{n}} \sum_{1 \le i \le n} Z_i U_i \xrightarrow{d} N(0, \text{Var}[ZU]) \ .$$

The desired result thus follows from the CMT.

8.1.3 Estimation of $\mathbb{V}$

A natural estimator of $\mathbb{V}$ is given by

$$\begin{aligned} \hat{\mathbb{V}}_n \quad = \quad & \left(\hat{\Pi}_n' \left(\frac{1}{n} \sum_{1 \le i \le n} Z_i Z_i' \right) \hat{\Pi}_n \right)^{-1} \\ & \times \hat{\Pi}_n' \left(\frac{1}{n} \sum_{1 \le i \le n} Z_i Z_i' \hat{U}_i^2 \right) \hat{\Pi}_n \\ & \times \left(\hat{\Pi}_n' \left(\frac{1}{n} \sum_{1 \le i \le n} Z_i Z_i' \right) \hat{\Pi}_n \right)^{-1} \ , \end{aligned}$$

where $\hat{U}_i = Y_i - X_i'\hat{\beta}_n$. As in our discussion of OLS, the primary difficulty in establishing the consistency of this estimator lies in showing that

$$\frac{1}{n}\sum_{1\leq i\leq n} Z_i Z_i' \hat{U}_i^2 \xrightarrow{P} \mathrm{Var}[ZU]$$

as $n \to \infty$. Note that part of the complication lies in the fact that we do not observe U_i and therefore have to use $\hat{U}_i$. However, the desired result can be shown by arguing exactly as in 480-2.

Note that $\hat{U}_i = Y_i - X_i'\hat{\beta}_n \neq Y_i - \hat{X}_i'\hat{\beta}_n$, so the standard errors from two repeated applications of OLS will be incorrect. Assuming $\mathrm{Var}[ZU]$ is invertible, inference may now be carried out exactly the same way as discussed for the OLS estimator, simply replacing the OLS quantities with their TSLS counterparts.

8.2 Efficiency of the TSLS Estimator

Let (Y, X, Z, U) be a random vector where Y and U take values in $\mathbf{R}$, X takes values in $\mathbf{R}^{k+1}$, and Z takes values in $\mathbf{R}^{\ell+1}$. Let $\beta \in \mathbf{R}^{k+1}$ be such that

$$Y = X'\beta + U \ .$$

Suppose $E[ZX'] < \infty$, $E[ZZ'] < \infty$, $E[ZU] = 0$, there is no perfect collinearity in Z, and that the rank of $E[ZX']$ is $k+1$. Denote by P the marginal distribution of (Y, X, Z). Let $(Y_1, X_1, Z_1), \ldots, (Y_n, X_n, Z_n)$ be an i.i.d. sequence of random variables with distribution P.

The TSLS estimator identifies β by means of the projection matrix $\Pi = E[ZZ']^{-1}E[ZX']$. However, note that we could have solved for β using any $(\ell+1)\times(k+1)$ dimensional matrix Γ such that $E[\Gamma' ZX']$ has rank $k+1$. For any such matrix,

$$\beta = E[\Gamma' ZX']^{-1}E[\Gamma' ZY] \ ,$$

and we could have estimated β using

$$\tilde{\beta}_n = \left(\frac{1}{n}\sum_{1\leq i\leq n} \Gamma' Z_i X_i'\right)^{-1}\left(\frac{1}{n}\sum_{1\leq i\leq n} \Gamma' Z_i Y_i\right) \ .$$

Note that one could use a consistent estimate of Γ, $\hat{\Gamma}_n$, instead. By arguing as before, it is possible to show under our assumptions that $\tilde{\beta}_n \xrightarrow{P} \beta$ as $n \to \infty$. If in addition $\mathrm{Var}[ZU] = E[ZZ'U^2] < \infty$, then, by arguing as before, it is also possible to show that

$$\sqrt{n}(\tilde{\beta}_n - \beta) \xrightarrow{d} N(0, \tilde{\mathbb{V}})$$

as $n \to \infty$, where

$$\tilde{\mathbb{V}} = E[\Gamma' Z X']^{-1} \Gamma' \mathrm{Var}[ZU] \Gamma E[\Gamma' Z X']^{-1\prime} .$$

We now argue that under certain assumptions, the "best" choice of Γ is given by Π, i.e., $\tilde{\mathbb{V}} \geq \mathbb{V}$.

In order to establish this claim, we assume that $E[U|Z] = 0$ and $\mathrm{Var}[U|Z] = \sigma^2$. In addition, define $W^* = \Pi' Z$ and $W = \Gamma' Z$, which will simplify the notation below. To see that $\tilde{\mathbb{V}} \geq \mathbb{V}$, first note that under these assumptions

$$\begin{aligned}
\tilde{\mathbb{V}} &= \sigma^2 E[\Gamma' Z X']^{-1} E[\Gamma' Z Z' \Gamma] E[\Gamma' Z X']^{-1\prime} \\
&= \sigma^2 E[\Gamma' Z Z' \Pi]^{-1} E[\Gamma' Z Z' \Gamma] E[\Gamma' Z Z' \Pi]^{-1\prime} \\
&= \sigma^2 E[W W^{*\prime}]^{-1} E[W W'] E[W W^{*\prime}]^{-1\prime}
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{V} &= \sigma^2 E[\Pi' Z Z' \Pi]^{-1} E[\Pi' Z Z' \Pi] E[\Pi' Z Z' \Pi]^{-1} \\
&= \sigma^2 E[\Pi' Z Z' \Pi]^{-1} \\
&= \sigma^2 E[W^* W^{*\prime}]^{-1} ,
\end{aligned}$$

where in both cases the first equality follows from $\mathrm{Var}[ZU] = E[ZZ'U^2] = \sigma^2 E[ZZ']$, and the second equality used the fact that $X = \Pi' Z + V$ with $E[ZV'] = 0$. It suffices to show that $\tilde{\mathbb{V}}^{-1} \leq \mathbb{V}^{-1}$, i.e., to show that

$$E[W^* W^{*\prime}] - E[W W^{*\prime}]' E[W W']^{-1} E[W W^{*\prime}] \geq 0 .$$

Yet this follows upon realizing that the left-hand side of the preceding display is simply $E[\hat{W}^* \hat{W}^{*\prime}]$ with

$$\hat{W}^* = W^* - \mathrm{BLP}(W^*|W) = W^* - E[W W^{*\prime}]' E[W W']^{-1} W .$$

When we do not assume that $E[U|Z] = 0$ and $\mathrm{Var}[U|Z] = \sigma^2$, then better estimators for β exist. Such estimators are most easily treated as a special case of the *generalized method of moments* (GMM), which we do not cover in this class.

8.3 TSLS for Binary Endogenous Variables

So far we considered the case where X takes values in $\mathbf{R}^{k+1}$ and Z takes values in $\mathbf{R}^{\ell+1}$. There are, however, a few important cases that arise frequently in applications and deserve special consideration. One of them is the case of a binary treatment variable, such as whether an individual attends college, has

an additional child, or receives some policy intervention. One of them is the case where $X = (1, A)'$, so that the model contains a constant and a binary variable denoted by A. In this case, the model is given by

$$Y = \beta_0 + \beta_1 A + U \ ,$$

where A is endogenous and so $E[AU] \neq 0$. This model, unlike OLS in general, is always intended to be causal and is never descriptive. The model implicitly states that the effect of A on Y is *homogeneous*, that is, every single individual reacts to A in exactly the same way and such an effect is captured by β_1. Of course this is a highly unattractive feature of the model, but later on in class we will discuss how the same model sometimes admits interesting interpretations even in cases with heterogeneous effects.

The case with a binary endogenous variable is quite popular in applications; we will see one such example in the next section. In these cases it is also often the case that the instrument, Z, is itself a binary random variable. When this happens, the IV estimand admits a particularly simple representation. Using the formulas for subvectors of β from Section 7.2.1, and noting that the best linear predictor of any random variable on a constant is its mean, we obtain

$$\beta_1 = \frac{E[Z(Y - E[Y])]}{E[Z(A - E[A])]} = \frac{\text{Cov}(Y, Z)}{\text{Cov}(A, Z)} \ . \tag{8.1}$$

However, since both A and Z are binary random variables, this formula can be further simplified. Dividing the numerator and denominator by $\text{Var}[Z]$ and using the fact that

$$\frac{\text{Cov}(V, Z)}{\text{Var}[Z]} = E[V \mid Z = 1] - E[V \mid Z = 0] \ , \tag{8.2}$$

for any random variable V such that $E[V] < \infty$ (see Problem 8.1), then it follows that

$$\beta_1 = \frac{E[Y \mid Z = 1] - E[Y \mid Z = 0]}{E[A \mid Z = 1] - E[A \mid Z = 0]} \ . \tag{8.3}$$

Thus, the IV estimand is the ratio of the reduced-form effect of Z on Y to the first-stage effect of Z on A. The right-hand side of (8.3) is called the Wald estimand (due to Wald [1940]), and so when A and Z are binary random variables, the IV estimand and Wald estimand coincide. Naturally, the sample analog of these quantities (i.e., the estimators of β_1) also coincide. The formula also illustrates what it means for the instrument to be "relevant" in this case; which is essentially the condition that guarantees that the denominator is non-zero. In order for this to be the case, we need that $E[A \mid Z = 1] \neq E[A \mid Z = 0]$, which in the case where A is binary can be written as,

$$P\{A = 1 \mid Z = 1\} \neq P\{A = 1 \mid Z = 0\} \ .$$

In other words, the instrument must affect the treatment or decision A, so that the probability of receiving treatment varies with the value of the instrument. Relevance alone is not enough, however: to interpret the Wald estimand causally, we also need the instrument to be exogenous.

8.4 Empirical Applications

8.4.1 Angrist and Evans

Angrist and Evans Angrist and Evans [1998] is one of the well-known papers of modern labor economics that represents a good example of endogenous decisions and valid instruments. The paper looks at the effect of fertility decisions (having an extra child) on female labor force participation and earnings. Without putting much thought into it, you could imagine tackling this question by running a regression of a woman's labor force status on whether or not she has a child or the number of children that she has. The correlation between the number of children and labor force participation tends to be strongly negative. But hopefully an instant later you would realize that the decision of having children is likely to be correlated with expected income. Indeed, these are joint *choices* that lead to many possible endogeneity stories. Here is just one: high-earning women have fewer children due to a higher opportunity cost.

To map the empirical strategy in Angrist and Evans to our notation, consider the following definitions. First, we denote by Y the outcome of interest, which will be a measure of labor market participation or outcome for the woman (or her husband). Angrist and Evans restrict the sample to only women (or couples) with 2 or more children, and so A is an indicator for having more than 2 children (versus exactly 2). That is,

$$A := I\{\text{number of children} > 2\} .$$

The authors then consider two instruments for A. The first one is $Z = 1$ if the first two children had the *same sex*. This is based on the idea that there is *preference to have a mix of boys and girls*, so that

$$P\{A = 1 \mid Z = 1\} > P\{A = 1 \mid Z = 0\} .$$

The instrument is also expected to be exogenous, as sex is "randomly" determined. The authors also consider another instrument, where $Z = 1$ if the second birth was a twin (i.e., the first two children are twins). The twins instrument is used primarily for comparison as it had been used in previous papers. While one could tell a history of how the probability of $A = 1$ may vary with such an instrument, exogeneity of the instrument is more difficult to defend as it is well-known that older women are more likely to have twins, on top of the fact that such an instrument affects both the number and spacing of children. Table 8.1 reproduces Table 4 in Angrist and Evans [1998] and shows that while same-sex is uncorrelated with a variety of observed confounders, the twins instrument is well-known to be correlated with age (so, education) and race.

Angrist and Evans compute the Wald estimators for a variety of outcomes, including labor income (which we denote by Y) below. If we let $N_1 = \{i \in$

TABLE 4—DIFFERENCES IN MEANS FOR DEMOGRAPHIC VARIABLES BY *SAME SEX* AND *TWINS*-2

	Difference in means (standard error)		
	By *Same sex*		By *Twins-2*
Variable	1980 PUMS	1990 PUMS	1980 PUMS
Age	−0.0147 (0.0112)	0.0174 (0.0112)	0.2505 (0.0607)
Age at first birth	0.0162 (0.0094)	−0.0074 (0.0114)	0.2233 (0.0510)
Black	0.0003 (0.0010)	0.0021 (0.0011)	0.0300 (0.0056)
White	0.0003 (0.0012)	−0.0006 (0.0013)	−0.0210 (0.0066)
Other race	−0.0006 (0.0005)	−0.0014 (0.0009)	−0.0090 (0.0041)
Hispanic	−0.0014 (0.0009)	−0.0007 (0.0010)	−0.0069 (0.0047)
Years of education	−0.0028 (0.0076)	0.0100 (0.0074)	0.0940 (0.0415)

Notes: The samples are the same as in Table 2. Standard errors are reported in parentheses.

TABLE 8.1: Differences in means for demographic variables by same-sex and twins instrument.

$\mathbf{N} : Z_i = 1\}$ and $N_0 = \{i \in \mathbf{N} : Z_i = 0\}$, then the numerator of the Wald estimator is given by

$$\frac{1}{|N_1|}\sum_{i=1}^{n} Y_i Z_i - \frac{1}{|N_0|}\sum_{i=1}^{n} Y_i(1-Z_i) = -132.5 \quad (34.4)\ ,$$

with standard errors between parenthesis, while the denominator is

$$\frac{1}{|N_1|}\sum_{i=1}^{n} A_i Z_i - \frac{1}{|N_0|}\sum_{i=1}^{n} A_i(1-Z_i) = 0.060 \quad (0.0016)\ .$$

Taking the ratio of the two yields

$$\hat{\beta}_1 = -2208.8(569.2)\ .$$

These numbers can be found in Table 5 of Angrist and Evans [1998]. The results show that, in addition to having more children than women with one

boy and one girl, women with two children of the same sex have lower annual earnings. The results in the paper also show that they have lower probability of working, work fewer weeks per year, and fewer hours per week.

8.4.2 Baum-Snow

Baum-Snow [2007] studies a classic question in urban economics: What explains the growth of suburbs in U.S. cities? One explanation, grounded in land use theory, is that improvements in transportation—specifically, faster commuting times—make suburban living more attractive. Other factors include shifting amenities, racial preferences, school desegregation, and crime in central cities.

To isolate the causal effect of transportation infrastructure, Baum-Snow examines the impact of highway construction between 1950 and 1990. He finds that roughly one-third of the suburbanization observed in this period can be attributed to the expansion of the highway system. Estimating this effect raises a natural endogeneity concern: are highways built because cities grow, or do cities grow because highways are built? For example, economically successful cities may build more highways to accommodate growth, or rising crime may spur suburban demand and justify road expansion.

TABLE 8.2: First Stage Results

	Dependent variable:
	Change in N. Highways, 1950 - 1990
Planed highways in 1950	0.510***
	(0.074)
1950 Center City Radius	0.306***
	(0.072)
Change in (log) income	−0.939
	(1.819)
Change in Metropolitan Area Population	0.856***
	(0.279)
Constant	0.463
	(1.231)
Observations	139
R^2	0.503
Note:	*p<0.1; **p<0.05; ***p<0.01

The identification strategy in Baum-Snow [2007] exploits the 1947 Federal Highway Plan, which laid out a national system of interstate highways to connect major cities. Crucially, local cities had little say in the planned routes. As a result, cities that were "assigned" more connections in the 1947 plan were effectively required to build more highway infrastructure, even if they had not

chosen to do so on their own. This provides a source of exogenous variation in highway construction. The author then runs the following regression:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \sum_{j=2}^{k} \beta_j X_{j,i} + U_i \ ,$$

where Y_i is the change in central-city population in city i between 1950 and 1990, $X_{1,i}$ is the change in the number of highways over the same period, and $X_{j,i}$ are control variables such as changes in income, initial population, and other demographic trends. Because cities may build highways in response to expected growth, the regressor $X_{1,i}$ is believed to be endogenous, and so $E[X_{1,i}U_i] \neq 0$.

To address this concern, the author uses as an instrument $Z_{1,i}$: the number of interstate highways assigned to city i in the 1947 plan. The first-stage regression is given by:

$$X_{1,i} = \pi_0 + \pi_1 Z_{1,i} + \sum_{j=2}^{k} \pi_j X_{j,i} + V_i \ .$$

Table 8.2 shows the first-stage results. The instrument appears to be *relevant*: conditional on controls, $Z_{1,i}$ is strongly positively correlated with $X_{1,i}$. Table 8.3 reports OLS and IV results. The LS coefficient on X_1 underestimates the suburbanization related to highway construction. The IV estimate is larger in magnitude, and plausibly causal: more highways caused a strong decrease in urban center population. The full `R` code used to produce the tables in this section is available on Canvas.

TABLE 8.3: Regression Results

	Dependent variable:	
	Change pop. urban city center, 1950-1990	
	OLS	*Instrumental Variable (IV)*
	(1)	(2)
Change in N. Highways, 1950 - 1990	-0.059^{***}	-0.123^{***}
	(0.014)	(0.029)
1950 Center City Radius	0.080^{***}	0.113^{***}
	(0.014)	(0.020)
Change in (log) income	0.084	0.048
	(0.335)	(0.362)
Change in Metropolitan Area Population	0.363^{***}	0.424^{***}
	(0.053)	(0.062)
Planed highways in 1950	-0.640^{***}	-0.588^{**}
	(0.227)	(0.246)
Observations	139	139
R^2	0.395	0.296
Note:		*p<0.1; **p<0.05; ***p<0.01

8.5 "Weak" Instruments

It turns out that the normal approximation justified by the preceding results can be poor in finite samples, especially when the rank of $E[ZX']$ is "close" to being $< k+1$. As a result, hypothesis tests and confidence regions based off of this approximation can behave poorly in finite-samples as well. To gain some insight into this phenomenon in a more elementary way, suppose

$$\begin{aligned} Y_i &= X_i\beta + U_i \\ X_i &= Z_i\pi + V_i \ , \end{aligned}$$

where $Z_1, \ldots, Z_n$ are non-random, $(U_1, V_1), \ldots, (U_n, V_n)$ is a sequence of i.i.d. $N(0, \Sigma)$ random vectors. Suppose $\pi \neq 0$. Consider the estimator given by

$$\hat{\beta}_n = \frac{\frac{1}{n}\sum_{i=1}^n Z_i Y_i}{\frac{1}{n}\sum_{i=1}^n Z_i X_i} \ .$$

Note that

$$\sqrt{n}(\hat{\beta}_n - \beta) = \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^n Z_i U_i}{\left(\frac{1}{n}\sum_{i=1}^n Z_i^2\right)\pi + \frac{1}{n}\sum_{i=1}^n Z_i V_i} \ .$$

The finite-sample, joint distribution of the numerator and denominator is simply

$$N\left(\begin{array}{c} 0 \\ \bar{Z}_n^2\pi \end{array} , \left(\begin{array}{cc} \bar{Z}_n^2\sigma_U^2 & \frac{1}{\sqrt{n}}\bar{Z}_n^2\sigma_{U,V} \\ \frac{1}{\sqrt{n}}\bar{Z}_n^2\sigma_{U,V} & \frac{1}{n}\bar{Z}_n^2\sigma_V^2 \end{array} \right) \right) \ ,$$

where

$$\bar{Z}_n^2 = \frac{1}{n}\sum_{i=1}^n Z_i^2 \ .$$

This joint distribution completely determines the finite-sample distribution of $\sqrt{n}(\hat{\beta}_n - \beta)$. In particular, it is the ratio of two (correlated) normal random variables. If $\bar{Z}_n^2 \to \bar{Z}^2$ as $n \to \infty$, then it is straightforward to show that

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} N\left(0, \frac{\sigma_U^2}{\pi^2\bar{Z}^2}\right) \ .$$

This approximation effectively treats the denominator like a constant equal to its mean, so we would expect it to be "good" when the mean is "large", i.e.,

$$\bar{Z}_n^2\pi \gg \frac{1}{\sqrt{n}}\sqrt{\bar{Z}_n^2}\sigma_V \ .$$

When π is "small" , i.e., however, the approximation may be quite poor in finite-samples. Note in particular that $\pi \neq 0$ is not sufficient for the approximation to be good in finite-samples.

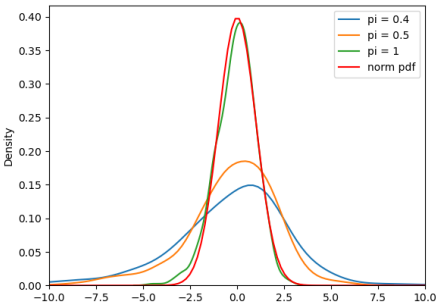


FIGURE 8.1: Empirical distribution of the IV estimate

Figure 8.1 shows the simulated distribution of $\sqrt{n}(\hat{\beta}_n - \beta)$ for different values of π. The sample size is $n = 50$, so we expect the weak IV problem to emerge when $\pi \gg \frac{1}{\sqrt{n}} \approx 0.14$ is violated. It can be seen from the figure that when $\pi = 0.5$, the empirical distribution of the IV estimate starts to deviate substantially from the normal distribution. Codes for replicating this figure and playing with different values of π are available.

A variety of ways of carrying out inference that do not suffer from this problem have been proposed in the literature. We now describe one simple and popular method. Consider the problem of testing the null hypothesis that $H_0 : \beta = c$ versus the alternative hypothesis $H_1 : \beta \neq c$ at level α. Note that under the null hypothesis, one can compute $U_i = Y_i - X_i'\beta$ and $Z_i U_i = Z_i(Y_i - X_i'\beta)$. Since it must be the case that $E[ZU] = 0$, we can simply test whether this is true using $Z_1 U_1, \ldots, Z_n U_n$. To formalize this idea, assume $\text{Var}[ZU]$ is invertible and define $W_i(c) = Z_i(Y_i - X_i'c)$. Note that when $\beta = c$, we have that

$$\sqrt{n}\bar{W}_n(c) = \frac{1}{\sqrt{n}} \sum_{1 \leq i \leq n} W_i(c) \xrightarrow{d} N(0, \Sigma(c)) \ ,$$

where $\Sigma(c) = \text{Var}[W(c)]$. Define

$$\hat{\Sigma}_n(c) = \frac{1}{n} \sum_{1 \leq i \leq n} (W_i(c) - \bar{W}_n(c))(W_i(c) - \bar{W}_n(c))' \ .$$

Using arguments given earlier, we see when $\beta = c$ that

$$T_n = n\bar{W}_n'(c)\hat{\Sigma}_n^{-1}(c)\bar{W}_n(c) \xrightarrow{d} \chi^2_{\ell+1} \ .$$

One may therefore test the null hypothesis by comparing T_n with $c_{\ell+1,1-\alpha}$, the $1-\alpha$ quantile of the $\chi^2_{\ell+1}$ distribution. One may now construct a confidence region using the duality between hypothesis testing and the construction of confidence regions. A closely related variant of this idea leads to the *Anderson-Rubin* test, in which one tests whether all of the coefficients in a regression of $Y_i - X_i'c$ on Z_i are zero.

Recent research in econometrics suggests that this method has good power properties when the model is exactly identified, but may be less desirable when the model is over-identified. Other methods for the case in which the model is over-identified and/or one is only interested in some feature of β (e.g., one of the slope parameters) have been proposed and are the subject of current research as well.

Instead of using these "more complicated" methods, researchers may attempt a two-step method as follows. In the first step, they would investigate whether the rank of $E[ZX']$ is "close" to being $< k+1$ or not by carrying out a hypothesis test of the null hypothesis that $H_0 : \text{rank}(E[ZX']) < k+1$ versus the alternative hypothesis that $H_1 : \text{rank}(E[ZX']) = k+1$. In some cases, such a test is relatively easy to carry out given what we have already learned: e.g., when there is a single endogenous regressor, such a test is equivalent to a test of the null hypothesis that certain coefficients in a linear regression are all equal to zero versus not all equal to zero. In the second step, they would only use these "more complicated" methods if they failed to reject this null hypothesis. This two-step method will also behave poorly in finite-samples and should not be used. A deeper discussion of these "uniformity" issues takes place in Econ 481.

8.6 Concluding Remarks

The material in this chapter borrows from notes by Azeem Shaikh that he kindly shared with me. In general, the topic of linear IV and endogeneity is covered in most standard sources, including Hansen [2022] and Wooldridge [2010].

8.7 Problems

Problem 8.1 *Prove* (8.2) .

Problem 8.2 *Consider the setting of Section 8.1.3. Show that if we use* $Y_i - \hat{X}_i'\hat{\beta}_n$ *instead of* $\hat{U}_i = Y_i - X_i'\hat{\beta}_n$*, the resulting estimator of* $\mathbb{V}$ *is inconsistent. You may assume* $k = \ell = 1$*.*

Problem 8.3 *Consider the linear model* $Y = X'\beta + U$*, and suppose you are not sure whether* X *is exogenous. You then proceed to test the null* H_0 :

$E[UX] = 0$ *by using the test statistic*

$$T_n = \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i\hat{U}_i\right)' \hat{\Sigma}^{-1} \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i\hat{U}_i\right),$$

where $\hat{\Sigma} \xrightarrow{p} E\left[XX'U^2\right]$. *The test you use is*

$$\phi_n = I\left\{T_n > c_{1-\alpha}\right\},$$

where $c_{1-\alpha}$ *is the* $(1-\alpha)$ *quantile of the* χ^2_k *distribution. Is this test asymptotically level* α*? What is the asymptotic power of the test against the alternative* $H_1 : E[XU] \neq 0$?

Problem 8.4 *Suppose* $X = (1, A)'$, *where* $A \in \{0, 1\}$ *is endogenous, and suppose the instrument* $Z \in \{0, 1\}$ *is binary and relevant. Consider the model*

$$Y = \beta_0 + \beta_1 A + U \ ,$$

with $E[ZU] = 0$ *and* $\mathrm{Cov}(A, Z) \neq 0$.

(a) Show that if $Y^* = Y + c$ *for some constant* $c \in \mathbf{R}$, *then the Wald estimand for* Y^* *is equal to the Wald estimand for* Y.

(b) Show that if $Y^* = aY + b$ *for constants* $a, b \in \mathbf{R}$ *with* $a \neq 0$, *then the Wald estimand for* Y^* *is equal to* $a\beta_1$.

(c) Show that if $A^* = 1 - A$, *then the Wald estimand based on* A^* *is equal to* $-\beta_1$.

Problem 8.5 *Suppose* $A \in \{0, 1, 2\}$ *is endogenous and consider the model*

$$Y = \beta_0 + \beta_1 I\{A = 1\} + \beta_2 I\{A = 2\} + U \ .$$

Let $Z \in \{0, 1\}$ *be a binary instrument satisfying* $E[ZU] = 0$.

(a) Show that

$$E[Y \mid Z = 1] - E[Y \mid Z = 0] = \beta_1\Delta_1 + \beta_2\Delta_2 \ ,$$

and provide expressions for Δ_1 *and* Δ_2.

(b) Explain why the parameters β_1 *and* β_2 *are not both identified with a single binary instrument. What standard IV requirement fails in this setting?*

Bibliography

J. D. Angrist and W. N. Evans. Children and their parents' labor supply: Evidence from exogenous variation in family size. *American Economic Review*, 88(3):450–477, 1998.

J. D. Angrist and J.-S. Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2008.

N. Baum-Snow. Did highways cause suburbanization? *The Quarterly Journal of Economics*, 122(2):775–805, 2007. doi: 10.1162/qjec.122.2.775.

B. Hansen. *Econometrics*. Princeton University Press, 2022.

A. Wald. The fitting of straight lines if both variables are subject to error. *The annals of mathematical statistics*, 11(3):284–300, 1940.

J. M. Wooldridge. *Econometric analysis of cross section and panel data*. MIT press, 2010.

9

Heterogeneous and Endogenous Treatments

Up until this point, we have focused our attention on a linear model with homogeneous effects. In such models,

$$Y = X'\beta + U \ ,$$

and the effect of a change in X (say, from $X = x$ to $X = x'$) is the *same* for everybody, and captured by the constant β. In this context, we have studied the properties of TSLS and learned that, unless strong assumptions are imposed, this estimator is not even efficient. Despite possible inefficiencies, TSLS remains popular. One reason is that its estimand has a well-understood interpretation under (unobserved) heterogeneity (i.e., cases where the effect of a change in X on Y would be expected to be different for different people). The easiest way to allow for such heterogeneity, as we have done in previous classes, is to allow for β to be *random*. When β is random, we may absorb U into the intercept and simply write

$$Y = X'\beta \ .$$

Note that this means that when we work with a random sample where variables are indexed by i, we would write $Y_i = X_i'\beta_i$, which makes it explicit that every individual has a unique effect β_i.

9.1 A Simple Random Coefficients Model

To get a better appreciation of the challenges associated with models that exhibit both endogeneity and unobserved heterogeneity, consider a random coefficients model with one endogenous covariate X_1:

$$Y = \beta_0 + \beta_1 X_1 \ .$$

Let Z_1 be an instrumental variable such that

$$X_1 = \pi_0 + \pi_1 Z_1 \ .$$

As opposed to the model in Section 7.3.1, where $(\beta_0, \beta_1, \pi_0, \pi_1)$ are unknown constants, here we allow them to be random variables. This allows each unit in

the population to have heterogeneous responses. In particular, β_1 represents the unit-specific causal effect of X_1 on Y, while π_1 represents the unit-specific effect of the instrument on the endogenous regressor.

To isolate the main point in the simplest possible way, assume that Z_1 is independent of $(\beta_0, \beta_1, \pi_0, \pi_1)$. Even under this favorable assumption, the IV estimand need not recover an average causal effect with a simple interpretation. In this setting, the IV estimand for the slope coefficient can be written as

$$\frac{\text{Cov}[Y, Z_1]}{\text{Cov}[X_1, Z_1]} = E\left[\omega\beta_1\right] \quad \text{where} \quad \omega = \frac{\pi_1}{E[\pi_1]} \ . \tag{9.1}$$

When $(\beta_0, \beta_1, \pi_0, \pi_1)$ are constants, we know that the IV estimand equals β_1. However, when these quantities are random, the IV estimand becomes a weighted average of the causal effect of X_1 on Y, captured here by β_1. Units for whom the instrument has a larger effect on X_1 receive more weight in absolute value. This immediately raises two concerns. First, some weights may be negative, namely for those units with $\pi_1 \times E[\pi_1] < 0$. Second, when we think about common instruments used in practice—such as distance from college, quarter of birth through compulsory schooling laws, or tuition subsidies—the IV estimand over-weights units who are especially sensitive to those instruments. Such units need not be representative of the overall population, or of the population relevant for a different policy intervention.

There are two broad ways to proceed after the lesson in (9.1). One approach is to keep the IV estimand on the left-hand side and develop assumptions that deliver a more concrete interpretation of the right-hand side by simplifying the form of the weights. The local average treatment effect (LATE) is the main concept that comes out of this line of thinking. The other approach is to give up on linear IV and change the estimator so as to directly recover a well-defined function of β_1. The rationale is natural: linear IV estimators were not designed for models with heterogeneous treatment effects. This changes both the left- and right-hand sides of (9.1), and leads to a variety of alternative approaches, from the marginal treatment effect (MTE) framework to partial identification methods. We discuss all of these approaches next.

9.2 Wald Estimand, Heterogeneity, and LATE

We now study the properties of TSLS, and the Wald estimand in particular, in the presence of heterogeneous effects of the variables X on Y. In order to provide the cleanest possible exposition, assume that $k = 1$ and let $X = (1, A)'$ with A being a binary random variable taking values in $\{0, 1\}$. In this notation,

$$Y = \beta_0 + \beta_1 A \ .$$

In this case, we interpret β_0 as $Y(0)$ and β_1 as $Y(1) - Y(0)$, where $Y(1)$ and $Y(0)$ are *potential* or *counterfactual outcomes*. Using this notation, we may rewrite the equation as

$$Y = AY(1) + (1 - A)Y(0) \ .$$

The potential outcome $Y(0)$ is the value of the outcome that would have been observed if (possibly counter-to-fact) A were 0; the potential outcome $Y(1)$ is the value of the outcome that would have been observed if (possibly counter-to-fact) A were 1. Following the same terminology we have been using in previous lectures, the variable A is called the *treatment*, $Y(1) - Y(0)$ is called the *treatment effect*, and the quantity $E[Y(1) - Y(0)]$ is referred to as the *average treatment effect*.

If A were randomly assigned (e.g., by the flip of a coin, as in a randomized controlled trial), then

$$(Y(0), Y(1)) \perp\!\!\!\perp A \ .$$

In this case, under mild assumptions, the slope coefficient from OLS regression of Y on a constant and A yields a consistent estimate of the average treatment effect; see Section 2.3.2.

We generally expect A to depend on $(Y(1), Y(0))$. For example, in the application in Angrist and Evans [1998] we would expect that the probability of having another child may be decreasing in $Y(0)$.

In this case, OLS will not yield a consistent estimate of the average treatment effect. To proceed further, we therefore assume, as usual, that there is an instrument Z that also takes values in $\{0, 1\}$. We may thus consider the slope coefficient from TSLS regression of Y on A with Z as an instrument. The estimand in this case is

$$\frac{\text{Cov}[Y, Z]}{\text{Cov}[A, Z]} = \frac{E[Y \mid Z = 1] - E[Y \mid Z = 0]}{E[A \mid Z = 1] - E[A \mid Z = 0]} \ ,$$

where the equality follows by multiplying and dividing by $\text{Var}[Z]$ and using earlier results. Our goal is to express this quantity in terms of the treatment effect $Y(1) - Y(0)$ somehow. To this end, analogously to our equation for Y above, it is useful to also introduce a similar equation for A:

$$\begin{aligned} A &= ZA(1) + (1 - Z)A(0) \\ &= A(0) + (A(1) - A(0))Z \\ &= \pi_0 + \pi_1 Z \ , \end{aligned}$$

where $\pi_0 = A(0)$, $\pi_1 = A(1) - A(0)$, and $A(1)$ and $A(0)$ are *potential* or *counterfactual treatments* (rather than outcomes). We impose the following versions of instrument exogeneity and instrument relevance, respectively:

$$(Y(1), Y(0), A(1), A(0)) \perp\!\!\!\perp Z$$

and

$$P\{A(1) \neq A(0)\} = P\{\pi_1 \neq 0\} > 0 \ .$$

Note that the first part of the assumption basically states that Z is as good as randomly assigned. In addition, note that we are implicitly assuming that Z does not affect Y directly, i.e., potential outcomes take the form $Y(a)$ as opposed to $Y(a, z)$. This is the exclusion restriction in this setting. In the linear model with constant effects, the exclusion restriction is expressed by the omission of the instruments from the causal equation of interest and by requiring that $E[ZU] = 0$.

We further assume the following *monotonicity* (or perhaps better called *uniform monotonicity*) condition:

$$P\{A(1) \geq A(0)\} = P\{\pi_1 \geq 0\} = 1 \ .$$

The monotonicity assumption states that while the instrument may have no effect on some people, all those who are affected are affected *in the same way*.

Under these assumptions, note that

$$\begin{aligned} E[Y \mid Z = 1] - E[Y \mid Z = 0] &= E[Y(1)A(1) + Y(0)(1 - A(1)) \mid Z = 1] \\ &\quad - E[Y(1)A(0) + Y(0)(1 - A(0)) \mid Z = 0] \\ &= E[Y(1)A(1) + Y(0)(1 - A(1))] \\ &\quad - E[Y(1)A(0) + Y(0)(1 - A(0))] \\ &= E[(Y(1) - Y(0))(A(1) - A(0))] \\ &= E[Y(1) - Y(0)|A(1) > A(0)]P\{A(1) > A(0)\} \ , \end{aligned}$$

where the first equality follows from the equations for Y and A, the second equality follows from instrument exogeneity, and the fourth equality follows from the monotonicity assumption. Furthermore,

$$E[A \mid Z = 1] - E[A \mid Z = 0] = E[A(1) - A(0)] = P\{A(1) > A(0)\} \ .$$

Hence, the Wald estimand equals

$$E[Y(1) - Y(0)|A(1) > A(0)] \ ,$$

which is termed the *local average treatment effect* (LATE). It is the average treatment effect among the subpopulation of people for whom a change in the value of the instrument switched them from being non-treated to treated. We often refer to such a subpopulation as *compliers*.

In the binary-binary case, it is useful to classify individuals into latent subpopulations according to $(A(0), A(1))$. Under monotonicity, there are three such groups: *never-takers*, for whom $(A(0), A(1)) = (0, 0)$; *always-takers*, for whom $(A(0), A(1)) = (1, 1)$; and *compliers*, for whom $(A(0), A(1)) = (0, 1)$. The fourth logical possibility, *defiers*, for whom $(A(0), A(1)) = (1, 0)$, is ruled

out by monotonicity. This classification is useful because the proportions of these groups are identified from the first stage:

$$P\{\text{never-taker}\} = P\{A = 0 \mid Z = 1\} ,$$
$$P\{\text{always-taker}\} = P\{A = 1 \mid Z = 0\} ,$$

and

$$P\{\text{complier}\} = P\{A = 1 \mid Z = 1\} - P\{A = 1 \mid Z = 0\} .$$

Thus, the first stage not only measures instrument relevance, but also identifies the size of the complier group.

A few remarks are in order: First, it is important to understand that this result depends crucially on the monotonicity assumption. Second, it is important to understand that this quantity may or may not be of interest. Third, it is important to understand that a consequence of this calculation is that in a world with heterogeneity "different instruments estimate different parameters." Finally, this result also depends on the simplicity of the model. When covariates are present, the entire calculation breaks down as we will illustrate below.

9.2.1 The Importance of Monotonicity

To better understand the role of monotonicity, it is useful to consider what happens without it. In that case, we would have

$$\begin{aligned} E[Y \mid Z = 1] - E[Y \mid Z = 0] = {} & E[Y(1) - Y(0) | A(1) > A(0)]P\{A(1) > A(0)\} \\ & - E[Y(1) - Y(0) | A(1) < A(0)]P\{A(1) < A(0)\} . \end{aligned}$$

Thus, without monotonicity, the reduced-form effect combines treatment effects for *compliers* and *defiers* with opposite signs. We might therefore have a situation where treatment effects are positive for everyone (i.e., $Y(1) - Y(0) > 0$) yet the reduced form is zero because effects on compliers are canceled out by effects on defiers, i.e., those individuals for whom the instrument pushes them out of treatment ($A(1) = 0$ and $A(0) = 1$).

This issue does not arise in a constant effect model where $\beta = Y(1) - Y(0)$ is the same for everyone. In that case,

$$\begin{aligned} E[Y \mid Z = 1] - E[Y \mid Z = 0] &= \beta\{P\{A(1) > A(0)\} - P\{A(1) < A(0)\}\} \\ &= \beta E[A(1) - A(0)] , \end{aligned}$$

and so a zero reduced-form effect means either the first stage is zero or $\beta = 0$.

The monotonicity assumption is difficult to defend in *many* empirical settings. Yet, it is widely used and researchers often perceive the assumption to be mild. Why? It turns out that monotonicity does hold in randomized controlled experiments with one-sided compliance. To see this, consider the context of randomized trials, where treatment assignment is independent of

potential outcomes by design. The fact that treatment is randomly assigned does not mean that every unit assigned to treatment actually takes the treatment. In medical trials, for example, units may be given access to a new medication for a disease but decide not to take it at home. In this case, one can interpret treatment assignment as an "offer of treatment" Z (the instrument), and actual treatment received as the variable A. This is the case in experiments where participation is voluntary among those randomly assigned to treatment.

At the same time, it is often the case that no one in the control group has access to the experimental intervention. In other words, $A(0) = 0$ while $A(1) \in \{0, 1\}$. It immediately follows that $A(1) \geq A(0)$ a.s., and monotonicity automatically holds. Since the group that receives treatment is then a self-selected subset of those offered treatment, a comparison between those actually treated ($A = 1$) and the control group ($A = 0$) is misleading.

In the setting we just described, two alternatives are frequently used. The first one is a comparison between those who were *offered* treatment ($Z = 1$) and the control group ($Z = 0$). This comparison is indeed based on randomly assigned Z and identifies a parameter known as the *intention-to-treat effect*. The second one is to use treatment assignment as an instrument for treatment received. In this case, IV solves the compliance problem discussed above. Moreover, under one-sided compliance, those who receive treatment are precisely the compliers, and so LATE coincides with the *treatment effect on the treated*, i.e., $E[Y(1) - Y(0) \mid A = 1]$. Table 9.3, at the end of the chapter, lists some empirical papers with IVs in randomized experiments.

9.3 An Application: Angrist and Evans revisited

Let's go back to the application we discussed in Section 8.4.1. Angrist and Evans study the effect of fertility decisions—in particular, having an additional child—on female labor force participation and earnings. To recap the notation, let Y denote the outcome of interest, such as labor force participation or earnings for the woman (or her husband). The authors restrict the sample to women (or couples) with at least two children, and so A is an indicator for having more than two children, as opposed to exactly two. That is,

$$A := I\{\text{number of children} > 2\} .$$

The authors consider two instruments for A, but the main one is

$$Z = I\{\text{the first two children have the same sex}\} .$$

Part of the reason the authors prefer this instrument over the twins instrument is that, as we discussed earlier, the twins instrument is less likely to be

	1980 PUMS		
		Wald estimate using as covariate:	
Variable	Mean difference by *Same sex*	*More than 2 children*	*Number of children*
More than 2 children	0.0600 (0.0016)	—	—
Number of children	0.0765 (0.0026)	—	—
Worked for pay	−0.0080 (0.0016)	−0.133 (0.026)	−0.104 (0.021)
Weeks worked	−0.3826 (0.0709)	−6.38 (1.17)	−5.00 (0.92)
Hours/week	−0.3110 (0.0602)	−5.18 (1.00)	−4.07 (0.78)
Labor income	−132.5 (34.4)	−2208.8 (569.2)	−1732.4 (446.3)
ln(*Family income*)	−0.0018 (0.0041)	−0.029 (0.068)	−0.023 (0.054)

TABLE 9.1: Table 5 in AE98. Wald estimates for five different outcomes and two instruments.

exogenous. But what about monotonicity? Monotonicity requires that

$$P\{A(1) \geq A(0)\} = 1 \ .$$

Equivalently, there are no *defiers*, that is, no individuals such that

$$\{A(1) = 0, A(0) = 1\} \ .$$

In this application, monotonicity rules out families who would choose to have a third child when the first two children have different sexes, but would choose not to have a third child when the first two children have the same sex. In other words, it rules out families whose preferences are such that they specifically want two children of the same sex. This is a strong requirement. The intuition that people tend to have *a preference for a mix of boys and girls* does not imply that there are no people who would rather have two boys or two girls. Monotonicity therefore restricts preference heterogeneity in a way that may be unattractive in this context.

Table 9.1 reports Wald estimates, including the one we presented in Section 8.4.1. Table 9.2, by contrast, reports the coefficient $\beta_{2\text{sls}}$ in the regression

$$Y = \beta_0 + \beta_{2\text{sls}} A + \gamma' W + U \ , \tag{9.2}$$

using Z as an excluded instrument for A, and where W is a vector of demographic covariates that includes age, age at first birth, and indicators for first

child boy, second child boy, black, Hispanic, and other race. Thus, Table 9.1 is closer to the clean binary-instrument / binary-treatment Wald case, whereas Table 9.2 already moves to a TSLS specification with covariates.

TABLE 7—OLS AND 2SLS ESTIMATES OF LABOR-SUPPLY MODELS USING 1980 CENSUS DATA

	All women			Married women			Husbands of married women		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Estimation method	OLS	2SLS	2SLS	OLS	2SLS	2SLS	OLS	2SLS	2SLS
Instrument for *More than 2 children*	—	*Same sex*	*Two boys, Two girls*	—	*Same sex*	*Two boys, Two girls*	—	*Same sex*	*Two boys, Two girls*
Dependent variable:									
Worked for pay	−0.176 (0.002)	−0.120 (0.025)	−0.113 (0.025) [0.013]	−0.167 (0.002)	−0.120 (0.028)	−0.113 (0.028) [0.013]	−0.008 (0.001)	0.004 (0.009)	0.001 (0.008) [0.013]
Weeks worked	−8.97 (0.07)	−5.66 (1.11)	−5.37 (1.10) [0.017]	−8.05 (0.09)	−5.40 (1.20)	−5.16 (1.20) [0.071]	−0.82 (0.04)	0.59 (0.60)	0.45 (0.59) [0.030]
Hours/week	−6.66 (0.06)	−4.59 (0.95)	−4.37 (0.94) [0.030]	−6.02 (0.08)	−4.83 (1.02)	−4.61 (1.01) [0.049]	0.25 (0.05)	0.56 (0.70)	0.50 (0.69) [0.71]
Labor income	−3768.2 (35.4)	−1960.5 (541.5)	−1870.4 (538.5) [0.126]	−3165.7 (42.0)	−1344.8 (569.2)	−1321.2 (565.9) [0.703]	−1505.5 (103.5)	−1248.1 (1397.8)	−1382.3 (1388.9) (0.549)
ln(*Family income*)	−0.126 (0.004)	−0.038 (0.064)	−0.045 (0.064) [0.319]	−0.132 (0.004)	−0.051 (0.056)	−0.053 (0.056) [0.743]	—	—	—
ln(*Non-wife income*)	—	—	—	−0.053 (0.005)	0.023 (0.066)	0.016 (0.066) [0.297]	—	—	—

TABLE 9.2: Table 7 in AE98. Reports estimates of $\beta_{2\text{sls}}$ in eq. (9.2).

Several features of these tables are worth noting. First, the OLS estimates tend to differ substantially from their IV counterparts, which is consistent with endogeneity or selection. Second, the Wald estimates, which do not control for covariates, tend to be farther from the OLS estimates than the corresponding estimates of $\beta_{2\text{sls}}$. Third, the interpretation of $\beta_{2\text{sls}}$ is already less transparent than the interpretation of the simple Wald estimand, because the inclusion of covariates moves us away from the clean binary-binary setup. Finally, Table 9.1 also reports Wald estimates when the variable A is redefined as the *number of children*, which is no longer binary. This introduces yet another departure from the simple Wald framework. We discuss the implications of some of these generalizations in the next section.

9.4 LATE's Generality

The beauty of LATE is mostly confined to the binary-binary case, in the absence of covariates. In this section we briefly cover the challenges that arise when we consider more general cases, with a focus on two particularly common extensions.

9.4.1 Multivalued Instruments

Binary instruments are common, but so are multivalued ones. Extending the LATE findings to the multivalued case helps illuminate both the intuition of the binary case and the interpretation difficulties that start to arise with more complicated specifications. Suppose that Z takes $L+1$ values $\{z_0, z_1, \ldots, z_L\}$. The treatment A is still binary, so now we have $L+1$ potential treatment assignments, one for each value of the instrument, and need to adjust our exogeneity and monotonicity assumptions as follows. Exogeneity now requires

$$Z \perp\!\!\!\perp (A(z_0), A(z_1), \cdots, A(z_L), Y(0), Y(1)) \ ,$$

whereas monotonicity requires

$$P\{A(z_0) \leq A(z_1) \leq \cdots \leq A(z_L)\} = 1 \ .$$

The monotonicity assumption states that the values of the instrument can be ordered in terms of their effect on the decision to take treatment, and that this order is the same for everyone. If Z represents different ordered costs or incentives to take treatment, then this assumption is usually not objectionable.

Multiple instrument values produce multiple complier groups. You might purchase a product only if the price is less than \$10, whereas another student might purchase if the price is less than \$20. The content of the monotonicity assumption is that you would both also purchase at \$5. If we move in the opposite direction and increase prices, there may come a point at which an individual decides to never buy. Different thresholds of this sort define different complier groups.

With multiple instrument values there are many ways to construct a linear IV estimator. We could use Z itself as a scalar instrument, or we could include indicators for each value of Z (i.e., $I\{Z = z_\ell\}$) using TSLS, among many other possibilities. Each choice leads to a different linear IV estimand with a potentially different interpretation. A general formulation is to replace Z by a scalar function $\zeta(Z)$ and then use the resulting variable as an instrument. Following the steps in Problem 9.3, it is possible to show that

$$\frac{\text{Cov}[Y, \zeta(Z)]}{\text{Cov}[A, \zeta(Z)]} = \sum_{\ell=1}^{L} \omega_\ell \text{LATE}_\ell \ , \tag{9.3}$$

where

$$\text{LATE}_\ell := \frac{E[Y \mid Z = z_\ell] - E[Y \mid Z = z_{\ell-1}]}{E[A \mid Z = z_\ell] - E[A \mid Z = z_{\ell-1}]} \ ,$$

and, for $\pi(z) := P\{A = 1 \mid Z = z\}$,

$$\omega_\ell := \frac{(\pi(z_\ell) - \pi(z_{\ell-1}))\,\text{Cov}[\zeta(Z), I\{Z \geq z_\ell\}]}{\text{Cov}[A, \zeta(Z)]} \ .$$

It follows that the IV estimand is a weighted average of the LATEs for each

of the L complier groups. Since $\pi(z_\ell) - \pi(z_{\ell-1}) \geq 0$, the sign of the weights ω_ℓ is determined by the covariance between $\zeta(Z)$ and $I\{Z \geq z_\ell\}$. This covariance is non-negative if ζ is an increasing function of Z; see Problem 9.3. So using Z by itself still produces a non-negatively weighted average of LATEs, albeit one where the weights depend on the cardinal values of Z, which might not be so attractive. The more cited form of the result takes $\zeta(z) = \pi(z)$ to be the propensity score, so that Z only affects the weights through the propensity score and its marginal distribution.

In the end, and despite the good news that we have obtained a weighted average of LATEs, the interpretation of the IV estimand in this context is much more difficult to articulate. It now includes multiple subpopulations weighted according to statistical considerations rather than conceptual ones.

9.4.2 LATE with covariates

While much of the LATE terminology evokes explicit randomization, that is not essential to the idea. When Z is not experimentally randomized, it is often important to condition on covariates in order to justify exogeneity. To illustrate this case, consider a binary instrument together with a set of covariates captured by the vector W. Exogeneity and monotonicity now become conditional requirements:

$$Z \perp\!\!\!\perp (A(0), A(1), Y(0), Y(1)) \mid W$$

and

$$P\{A(1) \geq A(0) \mid W\} = 1 \; a.s. \; .$$

Repeating the same arguments we used before, but now conditional on $W = w$, we obtain a conditional LATE:

$$\text{LATE}(w) = E[Y(1) - Y(0) \mid A(1) > A(0), W = w] \; . \tag{9.4}$$

The fact that conditioning on $W = w$ produces a LATE conditional on W is no more surprising than saying that removing all the men from the data leaves you with a conclusion that is conditional on women.

Perhaps the more relevant question is whether the TSLS estimand from a regression of Y on a constant and $X = (A, W)$ using (Z, W) as instruments—precisely the specification in (9.2)—admits a clear interpretation. If we let $\beta_{2\text{sls}}$ be the coefficient on A in this regression, then it follows from Problem 7.4 that

$$\beta_{2\text{sls}} = \frac{\text{Cov}[Z^*, Y]}{\text{Cov}[Z^*, A]} \quad \text{where} \quad Z^* = Z - \text{BLP}(Z|W) \; . \tag{9.5}$$

Is $\beta_{2\text{sls}}$ at least a weighted average of conditional LATEs? The answer, as shown recently by Blandhol et al. [2024], is no. The authors show that $\beta_{2\text{sls}}$ depends not only on functions of the treatment effects, but also on the level

of Y, and so it pulls in observed outcomes from not just the compliers, but also the always- and never-takers. The fact that $\beta_{2\text{sls}}$ depends on potential outcome levels—and not just treatment effects—makes interpretation problematic. Another way to phrase the issue is that the usual linear IV specification with covariates implicitly relies on a particular functional approximation to $E[Z \mid W]$, namely the best linear predictor of Z on W. Problem 9.8 asks you to think more carefully about why this matters for interpreting the TSLS coefficient on A. Thus, the interpretation of $\beta_{2\text{sls}}$ is tied not only to the causal structure of the model, but also to the way in which the dependence of Z on W is modeled. This helps explain why the simple LATE interpretation breaks down outside special cases such as saturated specifications with discrete covariates.

The one special case in which we recover a weighted average of conditional LATEs is when W is discrete and the TSLS regression is saturated in W. To formally define this regression, let $I_w := I\{W = w\}$ denote the indicator for the covariate W being equal to w. This approach then runs a regression of Y on a constant and $X = (A, \{I_w\}_{w \in \mathcal{W}})$ using Z as the excluded instrument. If we let β_A be the coefficient on A in this regression, then it follows that

$$\beta_A = E\left[\frac{\text{Cov}[A, Z \mid W]}{E[\text{Cov}[A, Z \mid W]]} \text{LATE}(W)\right] . \tag{9.6}$$

The intuition is that a saturated specification is only using variation in Z within W bins. By contrast, an unsaturated linear specification uses variation across both Z and W simultaneously. Even if we had discrete covariates, however, the LATE interpretation above still leaves much to be desired. Different covariate groups are given weight according to the covariance between the instrument and treatment within their group. This weight depends on the marginal distribution of Z and so reflects the assignment process of the instrument, not just the behavioral responses to the instrument. While the weights are non-negative, it is difficult to see what type of counterfactual question one could answer with the resulting weighted average.

9.5 Concluding Remarks

The material today borrows from several useful sources, most notably the lecture notes kindly shared by Alex Torgovitsky and material in Angrist and Pischke [2008]. I particularly want to thank Alex for sharing his source notes with me.

TABLE 9.3: Instruments used in Randomized Experiments

Outcome Variable	Endogenous Variable	Instrumental Variable(s)	Reference
Earnings	Participation in job training program	Random assignment of admission to training program	Bloom et al. (1997)
Earnings	Participation in Job Corps program	Random assignment of admission to training program	Burghardt et al. (2001)
Achievement test scores	Enrollment in private school	Randomly selected offer of school voucher	Howell et al. (2000)
Achievement test scores	Class size	Random assignment to a small or normal-size class	Krueger (1999)
Achievement test scores	Hours of study	Random mailing of test preparation materials	Powers and Swinton (1984)
Birth weight	Maternal smoking	Random assignment of free smoker's counseling	Permutt and Hebel (1989)

9.6 Problems

Problem 9.1 *Prove Equation* (9.1) *exploiting that* Z_1 *is independent of* $(\beta_0, \beta_1, \pi_0, \pi_1)$.

Problem 9.2 *Consider a randomized experiment where everybody assigned to treatment receives the treatment (like a vaccine), but people who are not randomly chosen to receive the treatment may obtain it outside the context of the experiment (say, they could obtain the vaccine somewhere else). Would monotonicity hold in this setting?*

Problem 9.3 *This problem walks you through the derivation and analysis of* (9.3).

(a) Show that the numerator can be written as

$$\operatorname{Cov}\left[Y, \zeta\left(Z\right)\right] = \sum_{\ell=0}^{L} \mathbb{E}\left[Y \mid Z = z_\ell\right] \left(\zeta\left(z_\ell\right) - E\left[\zeta\left(Z\right)\right]\right) P\left[Z = z_\ell\right].$$

(b) Show that

$$E\left[Y \mid Z = z_\ell\right] = E\left[Y \mid Z = z_0\right] + \sum_{j=1}^{\ell} \left(\pi\left(z_j\right) - \pi\left(z_{j-1}\right)\right) \mathrm{LATE}_j.$$

(c) Show that $\text{Cov}[Y, \zeta(Z)]/\text{Cov}[A, \zeta(Z)]$ *equals*

$$\sum_{\ell=1}^{L} \left[\frac{(\pi(z_\ell) - \pi(z_{\ell-1})) \sum_{j=\ell}^{L} (\zeta(z_j) - E[\zeta(Z)]) P[Z = z_j]}{\text{Cov}[A, \zeta(Z)]} \right] \text{LATE}_\ell.$$

(d) Show that

$$\sum_{j=\ell}^{L} (\zeta(z_j) - E[\zeta(Z)]) P[Z = z_j] = \text{Cov}[\zeta(Z), I\{Z \geq z_\ell\}].$$

(e) Show that if ζ *is an increasing function (meaning* $\ell \leq j$ *implies* $\zeta(z_\ell) \leq \zeta(z_j)$*), then for all* z_ℓ

$$\text{Cov}[\zeta(Z), I\{Z \geq z_\ell\}] \geq 0.$$

Problem 9.4 *Consider the binary treatment, binary instrument case, and suppose that the instrument exogeneity assumption holds. Weaken the monotonicity assumption to be the statement that either* $P\{A(1) \geq A(0)\} = 1$ *or* $P\{A(1) \leq A(0)\} = 1$*. How does this change the interpretation of the Wald estimand? Do you think this result is important in practice?*

Problem 9.5 *Consider the binary treatment, binary instrument case under the instrument exogeneity assumption and monotonicity assumption. Show that the following quantities are identified for any* y*; that is, they can be expressed as functions of the distribution of* (Y, A, Z)*:*

- $P\{Y(1) \leq y \mid A(1) = A(0) = 1\}$.
- $P\{Y(0) \leq y \mid A(1) = A(0) = 0\}$.
- $P\{Y(1) \leq y \mid A(1) > A(0)\}$.
- $P\{Y(0) \leq y \mid A(1) > A(0)\}$.

Provide an intuitive explanation of these identification results.

Problem 9.6 *Prove* (9.6).

Problem 9.7 *Consider the binary treatment, binary instrument case under instrument exogeneity and monotonicity. Let* $G \in \{at, nt, cp\}$ *denote whether an individual is an always-taker, never-taker, or complier. Show that*

$$P\{G = nt\} = P\{A = 0 \mid Z = 1\}, \qquad P\{G = at\} = P\{A = 1 \mid Z = 0\},$$

and

$$P\{G = cp\} = P\{A = 1 \mid Z = 1\} - P\{A = 1 \mid Z = 0\}.$$

Provide an intuitive explanation of each equality.

Problem 9.8 *Consider the setting of Section 9.4 with binary A and Z, and covariates W. Let*

$$\tilde{Z} = Z - \text{BLP}(Z \mid W).$$

(a) Show that

$$\text{Cov}(\tilde{Z}, Y) = E[\text{Cov}(\tilde{Z}, Y \mid W)] + E[E[\tilde{Z} \mid W]E[Y \mid W]].$$

(b) Show that

$$\text{Cov}(\tilde{Z}, A) = E[\text{Cov}(\tilde{Z}, A \mid W)] + E[E[\tilde{Z} \mid W]E[A \mid W]].$$

(c) Explain why the condition

$$E[\tilde{Z} \mid W = w] = 0 \quad \textit{for all } w$$

is important for interpreting the TSLS coefficient on A as a weighted average of conditional LATEs.

(d) Show that

$$E[\tilde{Z} \mid W] = 0$$

if and only if

$$E[Z \mid W] = \text{BLP}(Z \mid W).$$

Interpret this condition.

Bibliography

J. D. Angrist and W. N. Evans. Children and their parents' labor supply: Evidence from exogenous variation in family size. *American Economic Review*, 88(3):450–477, 1998.

J. D. Angrist and J.-S. Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2008.

C. Blandhol, J. Bonney, M. Mogstad, and A. Torgovitsky. When is tsls actually late? Technical report, National Bureau of Economic Research, 2024.

10

Marginal Treatment Effects

In the previous chapter we developed assumptions under which the Wald estimand recovers a well-defined causal parameter—the local average treatment effect (LATE)—for the subpopulation of compliers. The LATE is both the strength and the limitation of the approach: it is identified without imposing parametric restrictions, but it is specific to a particular instrument and a particular complier group. Different instruments estimate different LATEs, and without further assumptions there is no way to connect these estimates to the ATE, the ATT, or any other parameter defined over a broader population.

In this chapter we pursue a different approach. Rather than asking what the IV estimand identifies, we start from a structural model of treatment selection and use it to characterize a rich family of treatment effect parameters. The key concept is the *marginal treatment effect* (MTE), introduced by Heckman and Vytlacil [2005]. The MTE provides a single, unified representation of heterogeneous treatment effects that nests the ATE, ATT, and LATE as special cases. It also makes explicit what assumptions are needed to identify parameters beyond the LATE, and at what cost.

10.1 Roy Models

The classic econometric approach to modeling selection and heterogeneity is with a *selection equation*. Consider an individual deciding between treatment and no treatment. Suppose that their (indirect) utility from choosing $A = a$ is given by

$$V(a) - \nu_a(Z),$$

where $V(a)$ is unobservable and $\nu_a(Z)$ is a function of the instrument Z (and possibly covariates W, which we suppress for simplicity). Individuals choose the treatment that maximizes utility, so

$$A = I\{V(1) - \nu_1(Z) \geq V(0) - \nu_0(Z)\},$$

which after defining $\nu(Z) := \nu_0(Z) - \nu_1(Z)$ and $V := V(0) - V(1)$ simplifies to

$$A = I\{V \leq \nu(Z)\}. \tag{10.1}$$

Equation (10.1) is the selection equation. It is separable in the observable component $\nu(Z)$ and the unobservable component V. The variable V is assumed to be continuously distributed, a natural restriction given the latent utility motivation of the choice model. Without loss of generality we adopt the normalization $\nu(1) \geq \nu(0)$ (relabeling values of Z if needed).

The Roy model is completed with the outcome equation $Y = AY(1) + (1 - A)Y(0)$, the exclusion restriction $Y(z, a) := Y(a)$, and the following exogeneity condition:

$$Z \perp\!\!\!\perp (Y(0), Y(1), V). \tag{10.2}$$

The key feature of the Roy model is that it explicitly allows V to be correlated with $Y(0)$ and $Y(1)$. For example, $V(a)$ might be an individual's forecast of their lifetime earnings after attending college ($a = 1$) or not ($a = 0$). This forecast is likely imperfectly related to their actual earnings $Y(a)$ and it is also a determinant of the college choice. The Roy model captures this joint dependence in a tractable way.

10.2 Vytlacil's Equivalence Theorem

The Roy model is stated in terms of a latent variable V, while the LATE framework was stated in terms of potential treatment choices $A(0)$ and $A(1)$. A natural question is whether these two formulations impose the same restrictions on the data. Vytlacil [2002] shows they are equivalent.

Vytlacil (2002) Equivalence Theorem

The Roy model, meaning the selection equation $A = I\{V \leq \nu(Z)\}$ with $(Y(0), Y(1), V) \perp\!\!\!\perp Z$, is equivalent to the LATE assumptions of strong exogeneity $(Y(0), Y(1), A(0), A(1)) \perp\!\!\!\perp Z$ and monotonicity $P\{A(1) \geq A(0)\} = 1$.

The direction Roy $\Rightarrow$ LATE follows by defining $A(z) := I\{V \leq \nu(z)\}$. Monotonicity holds because, since either $\nu(z) \leq \nu(z')$ or $\nu(z') \leq \nu(z)$ for any z, z', we get $A(z) \leq A(z')$ or vice versa almost surely; the normalization $\nu(1) \geq \nu(0)$ then gives $A(0) \leq A(1)$ a.s. Strong exogeneity holds because each $A(z)$ is a deterministic function of V, and $(Y(0), Y(1), V) \perp\!\!\!\perp Z$. The direction LATE $\Rightarrow$ Roy requires constructing V and ν from the potential treatment choices; we omit this argument here.

The practical implication is that the Roy model and the LATE framework impose the same restrictions on the observed distribution of (Y, A, Z). The difference is only in notation and interpretation. For many economists, the Roy model is sometimes preferred because it provides a concrete choice-theoretic

foundation in which the role of observable and unobservable heterogeneity in selection is made explicit. That is, it is tightly linked to the idea that units make choices based on their preferences and information.

10.3 Marginal Treatment Effects

10.3.1 The Normalization

Working directly with V is cumbersome because its distribution F_V is unknown. The propensity score provides a convenient normalization. Note that

$$\pi(z) := P\{A = 1 \mid Z = z\} = P\{V \leq \nu(z)\} = F_V(\nu(z)) \ . \tag{10.3}$$

Since V is continuously distributed, F_V is continuous. Continuity of F_V alone suffices for the probability integral transform: $U := F_V(V) \sim \mathrm{U}[0,1]$. Moreover, F_V *is* strictly increasing on the support of V, and since V lies in its support almost surely, the equivalence $V \leq \nu(z) \iff F_V(V) \leq F_V(\nu(z))$ holds a.s. Therefore

$$A = I\{V \leq \nu(Z)\} = I\{F_V(V) \leq F_V(\nu(Z))\} = I\{U \leq \pi(Z)\} \ . \tag{10.4}$$

Two remarks. First, $U \perp\!\!\!\perp Z$ because $V \perp\!\!\!\perp Z$ and $U = F_V(V)$ is a function of V alone. Second, this representation expresses A directly as a threshold crossing of the uniform U at the propensity score $\pi(Z)$, with no remaining unknowns. We can interpret U as a latent measure of *resistance* to treatment: low values of U correspond to low resistance (high propensity to take treatment), high values to high resistance.

When Z is binary, the compliance types from the previous chapter correspond to regions of U:

$$\begin{array}{ll} \text{always-takers:} & U \leq \pi(0) \\ \text{compliers:} & \pi(0) < U \leq \pi(1) \\ \text{never-takers:} & U > \pi(1). \end{array} \tag{10.5}$$

The Roy model thus provides a sense of *magnitude* that was absent in the three-group classification. Some always-takers have $U \ll \pi(0)$ (extreme always-takers) while others have U only slightly below $\pi(0)$. Some compliers are close to the always-taker boundary and others are close to the never-taker boundary. The selection equation organizes all of this heterogeneity on a single continuous dimension.

10.3.2 Definition and Interpretation

The *marginal treatment effect* (MTE) function is the average treatment effect conditional on the unobservable resistance U:

$$\text{MTE}(u) = E[Y(1) - Y(0) \mid U = u]. \tag{10.6}$$

If $\text{MTE}(u)$ is decreasing in u, then those most resistant to treatment also experience the lowest treatment effects on average—a pattern consistent with comparative advantage in treatment. If $\text{MTE}(u)$ is constant in u, there is no unobservable heterogeneity in treatment effects, and distinctions between different target parameters collapse, as we will formalize in the next section.

The MTE also admits a direct interpretation as a *marginal* LATE. To see this, consider a binary instrument $Z \in \{0, 1\}$ with $\pi(1) \geq \pi(0)$. By (10.5), compliers are those with $U \in (\pi(0), \pi(1)]$, so writing $\Delta = Y(1) - Y(0)$,

$$\begin{aligned}
\text{LATE} &= E[\Delta \mid \text{complier}] = E[\Delta \mid \pi(0) < U \leq \pi(1)] \\
&= \frac{E[\Delta\, I\{\pi(0) < U \leq \pi(1)\}]}{P\{\pi(0) < U \leq \pi(1)\}} \\
&= \frac{\int_{\pi(0)}^{\pi(1)} E[\Delta \mid U = u]\, du}{\pi(1) - \pi(0)} \\
&= \frac{1}{\pi(1) - \pi(0)} \int_{\pi(0)}^{\pi(1)} \text{MTE}(u)\, du,
\end{aligned} \tag{10.7}$$

where the third equality uses $U \sim \text{U}[0, 1]$ (so the density of U is 1 and $P\{\pi(0) < U \leq \pi(1)\} = \pi(1) - \pi(0)$), and the fourth substitutes the definition of the MTE.

More generally, for any two instrument values z, z' with $\pi(z) \geq \pi(z')$, the LATE for the corresponding complier group is the average MTE over $(\pi(z'), \pi(z)]$. As $\pi(z) \searrow \pi(z')$, this average converges to $\text{MTE}(\pi(z'))$. The MTE is thus the limiting ("infinitesimal") version of LATE.

10.4 Target Parameters as Weighted Averages of the MTE

A central insight of the MTE framework is that many parameters of interest admit the representation

$$\tau = E\left[\int_0^1 \text{MTE}(u)\, \omega(u, Z)\, du\right] \tag{10.8}$$

for some weight function $\omega(u, z)$. This unified representation has two benefits. It shows which parts of the MTE function are relevant for each parameter;

and it makes explicit the source of identification difficulties, since identifying τ requires knowing the MTE at those values of u where ω is non-zero.

10.4.1 Average Treatment Effect

Since $U \sim \mathrm{U}[0,1]$, the ATE is the unweighted average of the MTE:

$$\mathrm{ATE} = E[\Delta] = E\Big[E[\Delta \mid U]\Big] = \int_0^1 \mathrm{MTE}(u)\, du. \tag{10.9}$$

Identifying the ATE nonparametrically requires the MTE to be identified everywhere on $[0,1]$, which in turn requires the propensity score to have full support: $\{0,1\} \subseteq \operatorname{supp} \pi(Z)$. This is essentially equivalent to having random assignment for extreme instrument values—a very strong requirement.

10.4.2 Average Treatment Effect on the Treated

The ATT involves a self-selected subpopulation. To derive its MTE representation, write

$$\begin{aligned} E[\Delta \mid A=1] &= E\left[\frac{\Delta A}{P\{A=1\}}\right] = E\left[\frac{E[\Delta A \mid Z, U]}{P\{A=1\}}\right] \\ &= E\left[\frac{\mathrm{MTE}(U)\, I\{U \le \pi(Z)\}}{P\{A=1\}}\right] \\ &= \int_0^1 \mathrm{MTE}(u)\, \frac{P\{\pi(Z) \ge u\}}{P\{A=1\}}\, du, \end{aligned} \tag{10.10}$$

where $\Delta = Y(1) - Y(0)$, and the last step used the law of total expectation over U and Z and the fact that $U \perp\!\!\!\perp Z$. Defining

$$\omega_{\mathrm{ATT}}(u) = \frac{P\{\pi(Z) \ge u\}}{P\{A=1\}}\ ,$$

the weights are highest for small u (always-takers get the most weight), decline as u increases through complier territory, and are zero for never-takers ($u > \sup \operatorname{supp} \pi(Z)$). The ATT thus over-represents those most inclined toward treatment relative to the full population.

10.4.3 Average Treatment Effect on the Untreated

By an analogous argument (Problem 10.2), the ATU is

$$E[\Delta \mid A=0] = \int_0^1 \mathrm{MTE}(u)\, \omega_{\mathrm{ATU}}(u)\, du, \tag{10.11}$$

where $\omega_{\mathrm{ATU}}(u) = P\{\pi(Z) < u\}/P\{A=0\}$. The weights are now concentrated at high values of u, reflecting that the untreated population is predominantly composed of never-takers.

10.4.4 Local Average Treatment Effect

For a binary instrument, the compliers have $U \in (\pi(0), \pi(1)]$ by (10.5), and the LATE is

$$\text{LATE} = \int_0^1 \text{MTE}(u)\, \omega_{\text{LATE}}(u)\, du \tag{10.12}$$

where

$$\omega_{\text{LATE}}(u) = \frac{I\{\pi(0) < u \leq \pi(1)\}}{\pi(1) - \pi(0)} \ .$$

The LATE puts uniform weight on compliers and zero weight on always- and never-takers—an equally-weighted average over $(\pi(0), \pi(1)]$, consistent with the interpretation from the previous chapter.

10.5 Identification of the MTE

10.5.1 Nonparametric Identification and Its Limits

Heckman and Vytlacil (2005) show that the MTE is nonparametrically identified at any p in the interior of $\operatorname{supp} \pi(Z)$ via the *local IV* estimand:

$$\text{MTE}(p) = \frac{\partial}{\partial p} E[Y \mid \pi(Z) = p]. \tag{10.13}$$

More generally, without additional functional-form assumptions, the data identify averages of the MTE over intervals determined by the support of $\pi(Z)$; pointwise identification of $MTE(p)$ requires enough smooth variation in $\pi(Z)$ around p, which is why the derivative formula applies only at interior support points.

Two important implications follow. First, the MTE is only nonparametrically identified on the interior of the support of $\pi(Z)$. Without further assumptions, parameters requiring knowledge of the MTE outside this support—including the ATE in general, and the ATT or ATU when the support excludes 0 or 1—are only partially identified. Second, with a binary instrument, the data identify only the average of the MTE over the interval $(\pi(0), \pi(1)]$, namely the LATE; they do not nonparametrically identify the MTE curve pointwise, and therefore do not identify the ATE or ATT without additional assumptions.

These observations have a direct implication for the external validity of the LATE. Equation (10.12) shows that a LATE identifies the MTE only on the interval $(\pi(0), \pi(1)]$. If the MTE is non-constant—if there is genuine unobservable heterogeneity in treatment effects—then extrapolating the LATE to the full population or to the compliers of a different instrument requires additional assumptions. Different instruments producing different LATEs is thus direct evidence against constancy of the MTE.

Identification of the MTE

First write $E[Y \mid \pi(Z) = p]$ using the law of iterated expectations:

$$\begin{aligned} E[Y \mid \pi(Z) = p] &= E[Y \mid A = 1, \pi(Z) = p]\, p + E[Y \mid A = 0, \pi(Z) = p]\,(1 - p) \\ &= E[Y(1) \mid A = 1, \pi(Z) = p]\, p + E[Y(0) \mid A = 0, \pi(Z) = p]\,(1 - p). \end{aligned}$$

Using the selection equation $A = I\{U \leq \pi(Z)\}$ and the independence $Z \perp\!\!\!\perp (Y(0), Y(1), U)$, conditioning on $[A = 1, \pi(Z) = p]$ is equivalent to conditioning on $[U \leq p]$ for expectations of $Y(1)$, and conditioning on $[A = 0, \pi(Z) = p]$ is equivalent to conditioning on $[U > p]$ for expectations of $Y(0)$. Thus

$$E[Y \mid \pi(Z) = p] = E[Y(1) \mid U \leq p]\, p + E[Y(0) \mid U > p]\,(1 - p).$$

Since $U \sim \mathrm{U}[0, 1]$, writing expectations in terms of the conditional density of U gives

$$E[Y \mid \pi(Z) = p] = \int_0^p E[Y(1) \mid U = u]\, du + \int_p^1 E[Y(0) \mid U = u]\, du.$$

Differentiating both sides with respect to p yields (10.13).

10.5.2 Parametric Identification via the MTR

When the instrument has limited support, a commonly used approach in practice is to impose restrictions on the shape of the MTE function. Define the *marginal treatment response* (MTR) functions

$$m_a(u) := E[Y(a) \mid U = u], \quad a \in \{0, 1\},$$

so that $\mathrm{MTE}(u) = m_1(u) - m_0(u)$. Restricting the MTR functions allows the MTE to be identified and extrapolated beyond the support of $\pi(Z)$.

The simplest specification imposes linearity:

$$m_a(u) = \alpha_a + \beta_a u, \quad a \in \{0, 1\}, \tag{10.14}$$

which implies a linear MTE:

$$\mathrm{MTE}(u) = (\alpha_1 - \alpha_0) + (\beta_1 - \beta_0)\, u. \tag{10.15}$$

The parameters (α_a, β_a) are identified by regressing Y on the propensity score within each treatment arm. To see this, note that under (10.14),

$$\begin{aligned} E[Y \mid A = 1, \pi(Z) = p] &= E[Y(1) \mid U \leq p] \\ &= \int_0^p (\alpha_1 + \beta_1 u) \frac{1}{p}\, du \\ &= \alpha_1 + \frac{\beta_1}{2} p\ . \end{aligned} \tag{10.16}$$

where we used that $U \mid U \le p \sim \mathrm{U}[0,p]$, so $dF_{U|U\le p} = \frac{1}{p}du$. Similarly,

$$\begin{aligned} E[Y \mid A=0, \pi(Z)=p] &= E[Y(0) \mid U > p] \\ &= \int_p^1 (\alpha_0 + \beta_0 u)\frac{1}{1-p}\,du \\ &= \left(\alpha_0 + \frac{\beta_0}{2}\right) + \frac{\beta_0}{2}\,p\,. \end{aligned} \tag{10.17}$$

Regressing Y on $\pi(Z)$ among treated units identifies $(\alpha_1,\, \beta_1/2)$; regressing among untreated units identifies $(\alpha_0 + \beta_0/2,\, \beta_0/2)$. Crucially, each regression only requires two distinct values of $\pi(Z)$, so a binary instrument is sufficient. Once (α_a, β_a) are identified, any parameter of the form (10.8) can be computed. For example,

$$\mathrm{ATE} = \int_0^1 \mathrm{MTE}(u)\,du = (\alpha_1 - \alpha_0) + \frac{1}{2}(\beta_1 - \beta_0). \tag{10.18}$$

More generally, one can use polynomial specifications

$$m_a(u) = \sum_{k=0}^{K} \beta_{a,k}\, u^k$$

of degree K, but identification then requires at least $K+1$ distinct values of $\pi(Z)$—one equation per unknown parameter, from a system like (10.16).

10.6 Application: Brinch, Mogstad, and Wiswall (2017)

Brinch, Mogstad, and Wiswall (2017) [Brinch et al., 2017, BMW] study the quantity-quality (Q-Q) tradeoff: the effect of family size on children's educational attainment. Resource dilution predicts a negative effect while gains from household stability might produce a positive one. Using Norwegian administrative data ($N \approx$ 514,000 families), they set Y to be the firstborn child's years of schooling. The treatment A and instrument Z follow Angrist and Evans (1998): A is an indicator for having more than two children (vs. exactly two), and Z is either same-sex or twins. Covariates W include cohort and parental education at first birth. Be mindful of notation: BMW write $P(Z)$ for what is $\pi(W,Z)$ in our notation, and their Z_- (the excluded instrument) is our Z.

Table 10.1 reports OLS and IV results from the regression $Y = \alpha + \beta A + W'\gamma + U$. In every specification, they include the same vector of observables (W, Z_1, Z_2) in the first-stage equations. What they change is the instrument(s) excluded from the outcome equation. In column 1, BMW use $P(Z)$ as the instrument, constructing it from logit estimates. When excluding the same-sex

TABLE 3
OLS AND IV ESTIMATES

	$P(Z)$ as Instrument (1)		Z_- as Instrument (2)
IV:			
Same-sex instrument	−.208		.174
	(.105)		(.115)
Twins instrument	−.065		.050
	(.060)		(.062)
Both instruments	−.015		.076
	(.053)		(.055)
OLS		−.052	
		(.007)	

TABLE 10.1: Table 3 in BMW. OLS and IV estimates of the effect of family size on the firstborn child's years of schooling. In column 1, $P(Z)$ is estimated via a logit model. Difference in rows is what is *e*xcluded from the outcome equation.

instrument from the outcome equation, the IV estimate is −0.208. When instead excluding the twins instrument, the point estimate remains negative but smaller. When both instruments are excluded simultaneously, the estimate is close to zero. In column 2, BMW use the instruments Z_- directly: the same-sex-based IV is −0.174 while the twins-based IV is only −0.050. The IV estimates thus vary in both magnitude and sign with the choice of excluded instrument. Recall that by the results in Section 9.4.2, these IV estimands cannot be easily interpreted as LATEs. Under constant treatment effects all instruments should produce the same population parameter (up to sampling error), so this divergence is evidence that the MTE is non-constant and that different instruments are recovering different weighted averages of it. The same-sex and twins instruments shift fertility for different groups of compliers, and those groups appear to experience different average treatment effects.

BMW respond to this by imposing the linear MTR specification (10.14), which allows point identification of the full MTE curve with a binary instrument. They also consider a partially identified approach that we discuss in the next chapter. The key advantage is that once the MTE is identified, any target parameter—ATE, ATT, or any other weighted average—can be computed analytically.

Table 10.2 reports the estimates of the linear MTR specification. By construction, the linear MTE model replicates the standard LATE (see Problem 10.4). More importantly, the slope of the linear MTE is statistically different from zero at conventional significance levels. This is a formal test of constant treatment effects: if $\beta_1 - \beta_0 = 0$, the MTE is flat and the LATE equals the ATE. Since the slope is nonzero, we reject the external validity

TABLE 4
ESTIMATES OF LINEAR MTE MODEL AND LATE BASED ON SAME-SEX INSTRUMENT

	$p = .473$	$p = .531$	Intercept	Slope
	A. Estimates of Linear MTE Model and Its Components			
Linear MTE model:				
$\mu_1 + K_1(P) = E(Y_1 \mid U_D < p)$	12.086	12.131	11.720	$+.775p$
	(.008)	(.007)	(.095)	(.188)
$\mu_0 + K_0(P) = E(Y_0 \mid U_D > p)$	12.462	12.450	12.564	$-.216p$
	(.007)	(.008)	(.091)	(.181)
$\mu_1 + k_1(p) = E(Y_1 \mid U_D = p)$	12.453	12.542	11.720	$+1.550p$
	(.084)	(.105)	(.095)	(.376)
$\mu_0 + k_0(p) = E(Y_0 \mid U_D = p)$	12.576	12.551	12.780	$-.432p$
	(.101)	(.080)	(.272)	(.0362)
$\text{MTE}(p) = E(Y_1 - Y_0 \mid U_D = p)$	$-.123$	$-.008$	-1.006	$+1.981p$
	(.129)	(.130)	(.290)	(.529)
	B. LATE from IV and Linear MTE Model			
Instrumental variables:				
$[E(Y \mid \Pr(D) = .531) - E(Y \mid \Pr(D) = .473)]/(.531 - .473)$				$-.065$
				(.129)
Linear MTE model:				
$\int_{.471}^{.531} \text{MTE}(p) = \text{MTE}[(.531 + .471)/2]$				$-.065$
				(.129)

NOTE.—This table displays LATE and linear MTE estimates of family size on the educational attainment of firstborn children. Panel A reports estimates from the linear MTE model with same sex, first and second as the excluded instrument. Panel B reports estimates of LATE from the IV estimator and the linear MTE model, with same sex, first and second as the excluded instrument. We do not include any covariates in the MTE estimation or the IV estimation. Standard errors in parentheses are computed by nonparametric bootstrap with 100 bootstrap replications.

TABLE 10.2: Table 4 in BMW. Linear MTR/MTE estimates. U_D in their notation is our U.

of the LATE and confirm that different instruments are indeed recovering different local parameters.

To extrapolate to the ATE under the linear specification, we apply (10.18):

$$\text{ATE} = (\alpha_1 - \alpha_0) + \frac{1}{2}(\beta_1 - \beta_0) = -1.06 + \frac{1.982}{2} \approx -0.069.$$

This is somewhat larger in magnitude than the LATE of -0.065, which reflects the fact that the MTE is declining in u: those less inclined to treatment experience larger (more negative) effects, and the ATE integrates over all of them. Similarly, one can compute an ATT of approximately 0.048, though as noted below these extrapolations should be interpreted with care.

BMW do not put much weight on the linear extrapolations, and for good reason. Their nonparametric estimates—using richer polynomial MTR specifications combined with covariates—show that the MTE is not well-approximated by a linear function; it exhibits a U-shape in u. The linear model forces a particular shape that is unlikely to be correct. This points to a general tension in the MTE approach: the richer the parametric specification, the more convincing the model, but the more support points the instrument must provide. BMW also consider polynomials of higher degree and show that one can identify at most a polynomial of degree $|\text{supp}\,\pi| - 1$, confirming the

fundamental constraint that identification is bounded by the support of the propensity score.

10.7 Concluding Remarks

The material in this chapter draws primarily on the lecture notes kindly shared by Alex Torgovitsky, and on the papers by Heckman and Vytlacil [2005], Vytlacil [2002], and Brinch et al. [2017]. For students wishing to go further, Heckman and Vytlacil (2005) is the canonical reference for the MTE identification theory. BMW provides a clean empirical application of the parametric MTR approach.

10.8 Problems

Problem 10.1 *Roy models typically assume, without loss of generality, that $\nu(1) \geq \nu(0)$. Suppose additionally that $V \sim \mathrm{U}[0,1]$. Show that the monotonicity assumption from the LATE framework holds automatically in this setting.*

Problem 10.2 *Following similar steps to those in (10.10), derive an expression for the ATU as a weighted average of the MTE. Characterize the weight function $\omega_{\mathrm{ATU}}(u)$ and verify that it integrates to one.*

Problem 10.3 *Show that the LATE can be written in the form (10.8) and derive the weight function $\omega_{\mathrm{LATE}}(u)$. Verify that it integrates to one.*

Problem 10.4 *Show that under the linear MTR specification (10.14), the LATE implied by the MTE representation (10.12) coincides with the Wald estimand.*

Problem 10.5 *Consider the linear MTR specification (10.14). Show that the ATT can be written as*

$$\mathrm{ATT} = (\alpha_1 - \alpha_0) + \frac{\beta_1 - \beta_0}{2} \cdot \frac{E[\pi(Z)^2]}{P\{A=1\}}.$$

Bibliography

C. N. Brinch, M. Mogstad, and M. Wiswall. Beyond late with a discrete instrument. *Journal of Political Economy*, 125(4):985–1039, 2017.

J. J. Heckman and E. Vytlacil. Structural equations, treatment effects, and econometric policy evaluation 1. *Econometrica*, 73(3):669–738, 2005.

E. Vytlacil. Independence, monotonicity, and latent index models: An equivalence result. *Econometrica*, 70(1):331–341, 2002.

11

Partial Identification

The previous chapter showed that the LATE is nonparametrically identified under IV assumptions, and that moving beyond the LATE to parameters like the ATE or ATT generally requires additional restrictions. When those restrictions are unavailable, or when we are unwilling to impose them, a natural next step is to consider a *partial identification*approach: what can be learned about a parameter of interest from the data, under a given set of assumptions, without necessarily pinning it down to a single value?

This chapter develops the partial identification approach in the context of treatment effects. We begin with the foundational framework of Manski [1989], who showed that the identified set for average potential outcomes has a natural "worst-case" characterization under bounded outcomes alone. We then examine how additional assumptions—Monotone Treatment Selection and Monotone IV—can tighten these bounds without requiring the full strength of an exclusion restriction. We discuss how sharp bounds can be derived under standard IV assumptions, and we return to the application of Brinch, Mogstad, and Wiswall (BMW) Brinch et al. [2017] to see how these bounds behave in practice. The chapter closes with an overview of Mogstad, Santos, and Torgovitsky Mogstad et al. [2018], who show how partial identification can be pursued within the MTE framework of the previous chapter, providing a unified approach to learning about treatment effect parameters under weak assumptions.

11.1 Worst-Case Bounds

The partial identification approach starts from a simple observation: the distribution of observed data cannot by itself reveal the distribution of counterfactual outcomes, because for each unit only one potential outcome is ever observed. This is the "anatomy of the selection problem" in the language of Manski [1989]. Any informative statement about counterfactual means therefore requires assumptions, and the question is how much those assumptions contribute to identification.

To fix ideas, suppose $Y(a) \in [Y_{\text{l}}, Y_{\text{u}}]$ for $a \in \{0, 1\}$, where Y_{l} and Y_{u} are known finite bounds on the outcome. We make no assumption about how

treatment A relates to $(Y(0), Y(1))$. For concreteness, consider identification of $E[Y(1)]$. By the law of iterated expectations,

$$E[Y(1)] = E[Y(1) \mid A = 1]\, P\{A = 1\} + E[Y(1) \mid A = 0]\, P\{A = 0\}. \quad (11.1)$$

The first term is identified from data: $E[Y(1) \mid A = 1] = E[Y \mid A = 1]$ by consistency, and $P\{A = 1\}$ is observed. The second term involves the mean of $Y(1)$ among units that do not take treatment, which is never observed. The only restriction we have imposed on this quantity is that $Y(1) \in [Y_{\mathrm{l}}, Y_{\mathrm{u}}]$. Using the bound on $E[Y(1) \mid A = 0]$, we immediately obtain the *worst-case bounds*:

$$E[Y \mid A = 1]\pi + Y_{\mathrm{l}}(1 - \pi) \;\le\; E[Y(1)] \;\le\; E[Y \mid A = 1]\pi + Y_{\mathrm{u}}(1 - \pi), \quad (11.2)$$

where $\pi = P\{A = 1\}$. By a symmetric argument,

$$E[Y \mid A = 0](1 - \pi) + Y_{\mathrm{l}}\pi \;\le\; E[Y(0)] \;\le\; E[Y \mid A = 0](1 - \pi) + Y_{\mathrm{u}}\pi. \quad (11.3)$$

Combining (11.2) and (11.3), the identified set for the ATE $= E[Y(1)] - E[Y(0)]$ is the interval $[\underline{\Delta}, \overline{\Delta}]$ where

$$\underline{\Delta} = E[Y \mid A = 1]\pi - E[Y \mid A = 0](1 - \pi) + Y_{\mathrm{l}}(1 - \pi) - Y_{\mathrm{u}}\pi, \quad (11.4)$$

$$\overline{\Delta} = E[Y \mid A = 1]\pi - E[Y \mid A = 0](1 - \pi) + Y_{\mathrm{u}}(1 - \pi) - Y_{\mathrm{l}}\pi. \quad (11.5)$$

The width of this interval is $\overline{\Delta} - \underline{\Delta} = (Y_{\mathrm{u}} - Y_{\mathrm{l}})$, which does not depend on the distribution of treatment. The interval always contains zero. This is a useful reminder that without any assumptions on selection, the sign of the ATE—let alone its magnitude—cannot be determined from data alone. The identified set is not vacuous, however: it does rule out values of the ATE outside $[Y_{\mathrm{l}} - Y_{\mathrm{u}}, Y_{\mathrm{u}} - Y_{\mathrm{l}}]$.

More generally, one can derive bounds conditional on covariates W by replacing unconditional with conditional expectations throughout. This does not change the width of the identified set, but it can be informative about how treatment effects vary across subgroups.

11.1.1 Monotone Treatment Selection

Manski and Pepper (2000) Manski and Pepper [2000] consider assumptions that are weaker than the standard unconfoundedness requirement but that nonetheless tighten the worst-case bounds. The first is Monotone Treatment Selection (MTS). Consider a binary treatment A and suppose we are interested in $E[Y(1)]$. MTS states that individuals who choose treatment have weakly higher mean potential outcomes under treatment than those who do not:

$$E[Y(1) \mid A = 0] \le E[Y(1) \mid A = 1] = E[Y \mid A = 1]. \quad (11.6)$$

This replaces the unknown $E[Y(1) \mid A = 0]$ in the worst-case upper bound with $E[Y \mid A = 1]$, which is identified. The identified set for $E[Y(1)]$ under MTS therefore becomes

$$\big[E[Y \mid A = 1]\pi + Y_{\mathrm{l}}(1 - \pi),\ E[Y \mid A = 1]\big]. \tag{11.7}$$

This upper bound is considerably tighter than the worst-case bound when π is small, because $E[Y \mid A = 1] < E[Y \mid A = 1]\pi + Y_{\mathrm{u}}(1 - \pi)$ whenever $Y_{\mathrm{u}} > E[Y \mid A = 1]$.

To see when MTS is plausible, suppose that A indicates completion of a doctoral degree and $Y(1)$ is an individual's expected earnings with a doctorate. MTS says that those who obtain a doctorate have higher mean earnings potential under that scenario than those who do not. This is consistent with positive selection on ability or motivation, which are standard features of models of educational choice. The analogous assumption for $Y(0)$ would require that those who do not obtain a doctorate have *lower* mean earnings without a doctorate—which rules out the possibility that high-ability individuals who choose not to pursue a doctorate would earn highly in that counterfactual state.

11.1.2 Monotone IV

Manski and Pepper (2000) Manski and Pepper [2000] also study identification with an instrument Z. Before turning to their approach, it is useful to recall the standard IV assumption: Z affects treatment assignment but is mean-independent of potential outcomes, $E[Y(a) \mid Z] = E[Y(a)]$ for all a. Standard IV can tighten the worst-case bounds: for each value of $Z = z$ one obtains bounds as in (11.2)–(11.3) (with propensity score $\pi(z) = P\{A = 1 \mid Z = z\}$ replacing π), and then bounds on $E[Y(a)]$ are obtained by averaging. The lower (resp. upper) bound under standard IV is the weighted average of the lower (resp. upper) conditional bounds, and it is no wider than the worst-case bound.

Manski and Pepper consider a strictly weaker assumption. The *Monotone IV* (MIV) assumption states that, for a scalar instrument Z,

$$E[Y(a) \mid Z = z] \le E[Y(a) \mid Z = z'] \quad \text{for all } z \le z' \text{ and all } a. \tag{11.8}$$

MIV does not require that Z is excluded from the potential outcome equations, nor that Z is independent of $(Y(0), Y(1))$. Instead, it only requires that conditional mean potential outcomes are weakly increasing in the instrument. Standard mean independence is a special case in which the conditional means are constant (and hence flat) in Z; MIV allows them to be increasing but not decreasing.

The MIV assumption is weaker than standard IV in an important sense: it permits Z to be a direct cause of outcomes and to be correlated with unobservables, as long as both channels operate in the same monotone direction.

To illustrate, suppose Z measures ability and $Y(a)$ is wages. Using Z as a standard IV requires that units with different measured ability have the same expected wages in each treatment state—which is strong. Using Z as an MIV only requires that higher-ability individuals have weakly higher mean wages in each state. This is consistent with Z being both a direct determinant of wages and correlated with unobserved factors.

To derive bounds under MIV, fix z and use the LIE to write:

$$E[Y(1) \mid Z = z] = E[Y \mid A = 1, Z = z]\,\pi(z) + E[Y(1) \mid A = 0, Z = z]\,(1 - \pi(z)).$$

The second term is unknown but bounded by $[Y_{\mathrm{l}}, Y_{\mathrm{u}}]$, yielding conditional worst-case bounds:

$$\begin{aligned} E[Y \mid A = 1, Z = z]\,\pi(z) + Y_{\mathrm{l}}(1 - \pi(z)) & \\ \leq\ & E[Y(1) \mid Z = z] \\ \leq\ & E[Y \mid A = 1, Z = z]\,\pi(z) + Y_{\mathrm{u}}(1 - \pi(z)). \end{aligned}$$

MIV further constrains $E[Y(1) \mid Z = z]$: by (11.8), it is no smaller than $E[Y(1) \mid Z = z_1]$ for any $z_1 \leq z$, and no larger than $E[Y(1) \mid Z = z_2]$ for any $z_2 \geq z$. Combining, the identified set for $E[Y(1) \mid Z = z]$ is $[LB(z), UB(z)]$ where

$$LB(z) = \sup_{z_1 \leq z} \Big(E[Y \mid A = 1, Z = z_1]\,\pi(z_1) + Y_{\mathrm{l}}(1 - \pi(z_1)) \Big), \tag{11.9}$$

$$UB(z) = \inf_{z_2 \geq z} \Big(E[Y \mid A = 1, Z = z_2]\,\pi(z_2) + Y_{\mathrm{u}}(1 - \pi(z_2)) \Big). \tag{11.10}$$

These are the sharp bounds on $E[Y(1) \mid Z = z]$ under MIV alone (Proposition 1 in Manski and Pepper [2000]). Bounds on the unconditional mean

$$E[Y(1)] = \sum_{z \in \mathcal{Z}} P\{Z = z\} E[Y(1) \mid Z = z]$$

then follow from

$$LB = \sum_{z \in \mathcal{Z}} P\{Z = z\}\, LB(z) \ \leq\ E[Y(1)] \ \leq\ \sum_{z \in \mathcal{Z}} P\{Z = z\}\, UB(z) = UB.$$

The structure of the MIV bounds has a natural interpretation: the lower bound on $E[Y(1) \mid Z = z]$ takes the most pessimistic conditional lower bound among all $z_1 \leq z$, exploiting the fact that the true value must be at least as large as any bound from below. Similarly for the upper bound. The gains relative to worst-case bounds depend on how much variation exists in $\pi(z)$ across values of the instrument.

What MIV does (and does not) require

The MIV assumption neither implies nor is implied by standard IV conditions. In a linear outcome model

$$Y(a,z) = \beta(a) + \gamma(z) + U \ ,$$

standard IV requires exclusion ($\gamma(z) = 0$ for all z) and mean exogeneity ($E[U \mid Z = z] = 0$). MIV allows both to fail, requiring only that their combined effect is weakly increasing in z: specifically, that $\gamma(z) + E[U \mid Z = z]$ is weakly increasing. Notably, the MIV bounds also allow for irrelevant instruments: $\pi(z) = \pi$ for all z is not ruled out, since no ratio of the form $1/(\pi(z_2) - \pi(z_1))$ appears in the derivations.

11.2 Sharp Bounds under IV Assumptions

The worst-case and MIV bounds do not require an exclusion restriction. When an instrument Z satisfies the full Roy model conditions—exclusion, independence $Z \perp\!\!\!\perp (Y(0), Y(1), U)$, and monotonicity—considerably sharper bounds become available. We derive these bounds in two steps. First, we exploit the Roy model structure directly to obtain bounds that use the propensity score support. Second, we show that Manski's IV bounds—derived from mean independence alone—coincide with the Roy model bounds.

11.2.1 Roy model bounds

Let $\mathcal{P} \equiv \text{supp}(\pi(Z))$ and let $\bar{p} = \sup \mathcal{P}$ and $\underline{p} = \inf \mathcal{P}$. For any $p \in \mathcal{P}$ we can write

$$\begin{aligned} E[Y(1)] &= E[Y(1) \mid \pi(Z) = p] \\ &= E[YA \mid \pi(Z) = p] + E[Y(1)(1-A) \mid \pi(Z) = p], \end{aligned} \tag{11.11}$$

where the first term is point identified and the second is not. Using the Roy model relationships from the previous chapter, the first term can be written as

$$E[YA \mid \pi(Z) = p] = E[Y(1) \mid U \leq p]\, p = \int_0^p E[Y(1) \mid U = u]\, du, \tag{11.12}$$

while the second term satisfies

$$E[Y(1)(1-A) \mid \pi(Z) = p] = \int_p^1 E[Y(1) \mid U = u]\, du. \tag{11.13}$$

The quantity in (11.12) is identified, but (11.13) is not—it involves $E[Y(1) \mid U = u]$ for $u > p$, which we do not observe directly. Since we want to identify as much of (11.12) as possible, we should take p as large as possible. At $p = \bar{p}$, the identified part is maximized. Bounding the unidentified remainder by $[Y_{\text{l}}, Y_{\text{u}}]$, we obtain

$$\begin{aligned}\underline{B}_1 \equiv E[YA \mid \pi(Z) = \bar{p}] + (1 - \bar{p})Y_{\text{l}} \\ \leq\ E[Y(1)] \leq \\ E[YA \mid \pi(Z) = \bar{p}] + (1 - \bar{p})Y_{\text{u}} \equiv \overline{B}_1.\end{aligned}$$

By a symmetric argument, using $p = \underline{p}$ to maximize the identified portion of $E[Y(0)]$:

$$\begin{aligned}\underline{B}_0 \equiv E[Y(1 - A) \mid \pi(Z) = \underline{p}] + \underline{p}\, Y_{\text{l}} \\ \leq\ E[Y(0)] \leq \\ E[Y(1 - A) \mid \pi(Z) = \underline{p}] + \underline{p}\, Y_{\text{u}} \equiv \overline{B}_0.\end{aligned}$$

Combining, the sharp identified set for the ATE under the Roy model assumptions is $[\underline{B}, \overline{B}]$ where

$$\underline{B} \equiv E[YA \mid \pi(Z) = \bar{p}] - E[Y(1 - A) \mid \pi(Z) = \underline{p}] + (1 - \bar{p})Y_{\text{l}} - \underline{p}\, Y_{\text{u}}, \tag{11.14}$$

$$\overline{B} \equiv E[YA \mid \pi(Z) = \bar{p}] - E[Y(1 - A) \mid \pi(Z) = \underline{p}] + (1 - \bar{p})Y_{\text{u}} - \underline{p}\, Y_{\text{l}}. \tag{11.15}$$

The width of this interval is $(Y_{\text{u}} - Y_{\text{l}})(1 - \bar{p} + \underline{p})$. It shrinks as the support of $\pi(Z)$ widens: if $\bar{p} = 1$ and $\underline{p} = 0$, the interval collapses to a point and the ATE is nonparametrically identified.

11.2.2 Manski's IV bounds and their equivalence to the Roy model bounds

Under the weaker assumption that Z is mean-independent of potential outcomes, $E[Y(a) \mid Z] = E[Y(a)]$ for all a, Manski's IV bounds proceed as follows. For any $z \in \mathcal{Z}$,

$$\begin{aligned}E[Y(1)] &= E[Y(1) \mid Z = z] \\ &= E[Y \mid A = 1, Z = z]\, \pi(z) + E[Y(1) \mid A = 0, Z = z]\, (1 - \pi(z)).\end{aligned}$$

Bounding the unidentified term and taking the tightest bound across all z:

$$\underline{B}_1^M \equiv \sup_{z \in \mathcal{Z}} \big\{ E[Y \mid A = 1, Z = z]\, \pi(z) + Y_{\text{l}}(1 - \pi(z)) \big\} \ \leq\ E[Y(1)], \tag{11.16}$$

$$E[Y(1)] \leq \inf_{z \in \mathcal{Z}} \big\{ E[Y \mid A = 1, Z = z]\, \pi(z) + Y_{\text{u}}(1 - \pi(z)) \big\} \equiv \overline{B}_1^M. \tag{11.17}$$

An analogous construction yields bounds $[\underline{B}_0^M, \overline{B}_0^M]$ for $E[Y(0)]$, and combining gives a sharp outer set for the ATE. The sharp identified set for the ATE under Manski's IV assumptions is $[\underline{B}^M, \overline{B}^M]$ with

$$\begin{aligned}\underline{B}^M \equiv & \sup_{z\in\mathcal{Z}}\big\{E[Y \mid A=1, Z=z]\,\pi(z) + Y_{\mathrm{l}}(1-\pi(z))\big\} \\ & - \inf_{z\in\mathcal{Z}}\big\{E[Y \mid A=0, Z=z]\,(1-\pi(z)) + Y_{\mathrm{u}}\,\pi(z)\big\}, \qquad (11.18)\\ \overline{B}^M \equiv & \inf_{z\in\mathcal{Z}}\big\{E[Y \mid A=1, Z=z]\,\pi(z) + Y_{\mathrm{u}}(1-\pi(z))\big\} \\ & - \sup_{z\in\mathcal{Z}}\big\{E[Y \mid A=0, Z=z]\,(1-\pi(z)) + Y_{\mathrm{l}}\,\pi(z)\big\}. \qquad (11.19)\end{aligned}$$

These two derivations use different assumptions, yet they produce the same bounds. To see this for the upper bound on $E[Y(1)]$, note that under the Roy model,

$$\begin{aligned}\overline{B}_1^M &= \inf_{z\in\mathcal{Z}}\big\{E[Y \mid A=1, Z=z]\,\pi(z) + Y_{\mathrm{u}}(1-\pi(z))\big\} \\ &= \inf_{z\in\mathcal{Z}}\big\{E[Y(1) \mid U \le \pi(z)]\,\pi(z) + Y_{\mathrm{u}}(1-\pi(z))\big\} \\ &= \inf_{p\in\mathcal{P}}\big\{E[Y(1) \mid U \le p]\,p + Y_{\mathrm{u}}(1-p)\big\} \\ &\le E[Y(1) \mid U \le \bar{p}]\,\bar{p} + Y_{\mathrm{u}}(1-\bar{p}) = E[YA \mid \pi(Z)=\bar{p}] + Y_{\mathrm{u}}(1-\bar{p}) = \overline{B}_1^R.\end{aligned}$$

Since the Roy model imposes strictly stronger assumptions, we also have $\overline{B}_1^R \le \overline{B}_1^M$, so $\overline{B}_1^R = \overline{B}_1^M$. A symmetric argument establishes equality of the lower bounds. Hence the Roy model and Manski IV bounds for the ATE coincide: the additional structure of the Roy model does not tighten the bounds relative to mean independence alone.

The width of the identified set under IV assumptions depends directly on instrument strength. For a binary instrument $Z \in \{0,1\}$ with $\pi(1) > \pi(0)$, we have $\bar{p} = \pi(1)$ and $\underline{p} = \pi(0)$, and the ATE bounds have width $(Y_{\mathrm{u}} - Y_{\mathrm{l}})(1 - \pi(1) + \pi(0))$. A stronger instrument (larger $\pi(1) - \pi(0)$) yields strictly tighter bounds.

11.3 Application: Bounds in the Q-Q Model

Brinch, Mogstad, and Wiswall (2017) Brinch et al. [2017] compute worst-case bounds and sharp IV bounds for the ATE and ATT in their study of family size effects on children's education. As discussed in the previous chapter, the outcome Y is the firstborn child's years of education, treatment A is an indicator for having more than two children, and the instruments are the same-sex

and twins indicators used by Angrist and Evans (1998). Since educational attainment is bounded between 0 and some maximum ($Y_{\text{u}} \approx 20$ years), the bounded outcome restriction is natural.

The BMW Appendix E.1 reports the following results. Under worst-case bounds alone, the identified set for the ATE is approximately $[-7.7, 7.3]$ and for the ATT is approximately $[-8.9, 6.1]$. These intervals are uninformative: they span nearly 15 units of education in each direction, which is far larger than any plausible effect of family size.

Adding the full IV assumptions (exclusion, independence, and monotonicity) tightens the bounds somewhat. With the same-sex instrument, the ATE falls in approximately $[-7.2, 6.9]$ and the ATT in approximately $[-8.4, 5.7]$. With the twins instrument, the ATE is approximately $[-4.3, 3.1]$ and the ATT is approximately $[-8.8, 6.0]$. The twins instrument provides a tighter upper bound on the ATE because it induces a larger shift in the propensity score, so more of the $[0, 1]$ support of U is identified.

Why IV bounds are tighter than worst-case bounds

Under IV assumptions and monotonicity, the propensity score shift $\pi(z_2) - \pi(z_1) > 0$ identifies the MTE over the interval $(\pi(z_1), \pi(z_2))$ via Local IV. The identified set for the ATE (which integrates the MTE over all of $[0, 1]$) therefore shrinks as the support of $\pi(Z)$ grows. The twins instrument, which produces a larger shift in the propensity score than same-sex, covers a larger portion of $[0, 1]$ and thus delivers tighter bounds on the ATE. The ATT, however, places more weight on units with U near zero (those always likely to choose treatment), and this region may be equally poorly identified under both instruments—explaining why the ATT bounds are similarly wide regardless of the choice of instrument.

Taken together, the BMW bounds are quite wide. The authors conclude that nonparametric bounds alone are not informative about the sign, let alone the magnitude, of the family-size effect. This motivates the linear MTE parameterization of the previous chapter, which achieves point identification by imposing functional form restrictions on the MTR functions. The tension between those parametric assumptions and the uninformative nonparametric bounds is the central message: identification is bought at the cost of functional form restrictions whose validity may be difficult to assess.

11.4 Partial Identification in the MTE Framework

The approach of Mogstad, Santos, and Torgovitsky (2018) Mogstad et al. [2018], hereafter MST, provides a unified framework for partial identification

of treatment effect parameters within the MTE structure. It generalizes the IV bounds of the previous section in two important directions: it allows for a flexible set of assumptions on the MTR functions (beyond bounded outcomes), and it provides a computationally tractable method for finding sharp bounds on any target parameter expressible as a weighted integral of the MTE.

11.4.1 Setup and target parameters

Recall from the previous chapter that under the Roy model assumptions, any mean treatment effect parameter of interest can be written as

$$\beta^\star(m_0, m_1) = \sum_{d \in \{0,1\}} E\left[\int_0^1 m_d(u, X)\, \omega_d^\star(u, X, Z)\, du\right], \tag{11.20}$$

where $m_d(u) = E[Y(d) \mid U = u]$ is the marginal treatment response (MTR) function for treatment arm d, and $\omega_d^\star$ are weights that depend only on the target parameter and the joint distribution of (X, Z)—both of which are identified from data. The ATE, ATT, ATU, LATE, and policy-relevant treatment effects all take this form for appropriate choices of $\omega_d^\star$.

The MST approach takes the target parameter $\beta^\star$ as the starting point—"forward engineering" the analysis rather than deriving parameters as a byproduct of what the instrument identifies. Identification is then a question about what the data and the maintained assumptions imply for the set of admissible MTR pairs (m_0, m_1), and in turn for the set of values $\beta^\star$ can take.

11.4.2 Observational constraints and IV-like estimands

The data impose constraints on the MTR functions through their moments. MST define an *IV-like estimand* as any quantity of the form

$$\beta_s \equiv E[Y \cdot s(A, X, Z)], \tag{11.21}$$

where s is a known function of the observables. For a given MTR pair (m_0, m_1), the value this estimand "would generate" is

$$\Gamma_s(m_0, m_1) = \sum_{d \in \{0,1\}} E\left[\int_0^1 m_d(u, X)\, \omega_{ds}(u, X, Z)\, du\right], \tag{11.22}$$

where the identified weights are $\omega_{1s}(u, x, z) = s(1, x, z)\, I\{u \le \pi(x, z)\}$ and $\omega_{0s}(u, x, z) = s(0, x, z)\, I\{u > \pi(x, z)\}$. Notice that Γ_s has exactly the same structure as $\beta^\star$: both are linear in (m_0, m_1).

A collection of IV-like estimands $\{\beta_s : s \in \mathcal{S}\}$ imposes the observational equivalence constraint $\Gamma_s(m_0, m_1) = \beta_s$ for all $s \in \mathcal{S}$. Any instrument-based moment—the Wald estimand, a regression of Y on A and Z, the LATE—can

be expressed in this form. By choosing $\mathcal{S}$ to include standard IV estimands, the MST bounds automatically reproduce the LATEs identified by the data, ensuring that the bounds are consistent with existing point-identified results.

11.4.3 Sharp bounds as a linear program

The identified set for $\beta^\star$ is determined by the optimization problem:

$$\overline{\beta}^\star = \sup_{(m_0,m_1)} \beta^\star(m_0, m_1) \quad \text{s.t.} \quad m \in \mathcal{M}, \quad \Gamma_s(m) = \beta_s \; \forall \; s \in \mathcal{S}, \tag{11.23}$$

and symmetrically for the lower bound $\underline{\beta}^\star$. Here $\mathcal{M}$ is the set of MTR functions satisfying the maintained assumptions. Crucially, since $\beta^\star$ and each Γ_s are linear in (m_0, m_1), this is a linear program if $\mathcal{M}$ is a polyhedral set. MST show that many economically relevant assumptions on the MTR functions do define polyhedral sets, making the bounds computationally tractable.

To operationalize the optimization, MST parameterize the MTR functions using a linear basis:

$$m_d(u, x) = \sum_{k=1}^{K_d} \theta_{dk} \, b_{dk}(u, x), \tag{11.24}$$

where b_{dk} are known basis functions and θ_{dk} are unknown coefficients. The optimization problem then becomes finite-dimensional and linear in θ. Two classes of basis functions are particularly useful:

- *Bernstein polynomials*: $b_k^K(u) = \binom{K}{k} u^k (1-u)^{K-k}$ for $k = 0, 1, \ldots, K$. These are well suited for imposing shape restrictions (monotonicity, concavity) via linear constraints on θ.
- *Constant splines*: $b_k(u) = I\{c_{k-1} \le u < c_k\}$ for a grid of knots c_k. When the knots coincide with the support points of $\pi(Z)$, constant splines exactly replicate the nonparametric bounds.

The following assumptions can be incorporated as linear constraints:

Boundedness. $Y \in [Y_\text{l}, Y_\text{u}]$ translates into box constraints on θ_{dk}.

Monotonicity. Assuming $m_0(u)$ is decreasing in u (positive selection bias) or $(m_1 - m_0)(u)$ is decreasing (decreasing returns to treatment) translates into linear inequalities on θ_{dk}.

Separability. Assuming $m_d(u, x) = m_d^U(u) + m_d^X(x)$ removes interaction terms and constrains the basis to be additively separable.

11.4.4 Sharpness and the role of instrument support

MST show that the bounds $[\underline{\beta}^\star, \overline{\beta}^\star]$ are sharp: every value in the interval corresponds to some MTR pair in $\mathcal{M}$ that is consistent with the data. The

bounds tighten as $\mathcal{S}$ grows (more IV-like estimands are included), with no loss beyond computational cost.

A key insight from this framework is that the width of the identified set has two components: how much of $[0,1]$ is covered by the support of $\pi(Z)$—the "intensive" variation in the instrument—and how much weight $\omega^\star$ places on the region outside the support, where MTR functions must be extrapolated. When the target parameter places substantial weight on regions far from the observed propensity scores, bounds will tend to be wide even with shape restrictions. The degree of extrapolation required is thus an explicit, measurable feature of the identification problem.

In a numerical illustration motivated by the bed-nets application of Dupas (2014), MST demonstrate that: (i) polynomial approximations of increasing degree converge to the nonparametric bounds; (ii) shape restrictions on the MTR functions—such as monotonicity—can have a large effect on the bounds; and (iii) the bounds for the ATT tend to be substantially wider than for extrapolated LATEs, because the ATT places weight on the region near $u = 0$ where few individuals are observed.

11.5 Concluding Remarks

The material in this chapter draws on Manski [1989], Manski and Pepper [2000], Heckman and Vytlacil [2005], Brinch et al. [2017], and Mogstad et al. [2018]. The book Manski [2003] provides an excellent general introduction to partial identification. I am grateful to Alex Torgovitsky for generously sharing his lecture notes.

11.6 Problems

Problem 11.1 *Let $Y \in [0,1]$ and consider identifying $E[Y(1)]$ without assumptions on the selection of A. Show that the worst-case identified set for $E[Y(1)]$ is $[E[Y \mid A=1]\,\pi,\ E[Y \mid A=1]\,\pi + (1-\pi)]$, where $\pi = P\{A=1\}$. Verify that this interval has width $1-\pi$ regardless of the treatment effect.*

Problem 11.2 *Derive the worst-case identified set for the $ATT = E[Y(1) - Y(0) \mid A=1]$ under bounded outcomes $Y \in [Y_{\mathrm{l}}, Y_{\mathrm{u}}]$ only. Show that the identified set always contains zero and compute its width.*

Problem 11.3 *Suppose the MTS assumption (11.6) holds and that $Y \in [Y_{\mathrm{l}}, Y_{\mathrm{u}}]$. Write down the identified set for the $ATE = E[Y(1)] - E[Y(0)]$ under*

MTS. Compare its width to the worst-case interval and discuss under what conditions MTS provides the largest improvement.

Problem 11.4 *Consider the sharp IV bounds for $E[Y(1)]$ derived in Section 11.2.2 for a binary instrument $Z \in \{0,1\}$ with $\pi(1) > \pi(0)$ and $Y \in [Y_{\mathrm{l}}, Y_{\mathrm{u}}]$.*

(a) Verify that the Manski IV upper bound (11.17) reduces to $E[Y \mid A = 1, Z = 1]\,\pi(1) + Y_{\mathrm{u}}(1 - \pi(1))$ for a binary instrument, and explain why $z = 1$ achieves the infimum.

(b) Show that the width of the sharp IV identified set for the ATE is $(Y_{\mathrm{u}} - Y_{\mathrm{l}})(1-\pi(1)+\pi(0))$. Verify that this equals the width formula derived from the Roy model bounds (11.14)–(11.15).

(c) Explain why a stronger instrument (larger $\pi(1)-\pi(0)$) tightens the bounds on the ATE but not necessarily on the ATT.

Problem 11.5 *This problem verifies directly why Manski's IV bounds coincide with the Roy model bounds in Section 11.2.2. Let $m_1(u) = E[Y(1) \mid U = u]$ and suppose $Y(1) \in [Y_{\mathrm{l}}, Y_{\mathrm{u}}]$. For $p \in \mathcal{P}$, define*

$$Q_u(p) = \int_0^p m_1(u)\,du + Y_{\mathrm{u}}(1-p), \qquad Q_l(p) = \int_0^p m_1(u)\,du + Y_{\mathrm{l}}(1-p)\ .$$

(a) Show that $Q_u(p)$ is weakly decreasing in p, while $Q_l(p)$ is weakly increasing in p.

(b) Conclude that, if $\bar{p} \in \mathcal{P}$, Manski's upper and lower IV bounds for $E[Y(1)]$ are attained at $p = \bar{p}$ and equal the Roy model bounds $\overline{B}_1$ and $\underline{B}_1$.

(c) Explain briefly why the analogous argument for $E[Y(0)]$ uses $\underline{p}$ rather than $\bar{p}$.

Bibliography

C. N. Brinch, M. Mogstad, and M. Wiswall. Beyond late with a discrete instrument. *Journal of Political Economy*, 125(4):985–1039, 2017.

J. J. Heckman and E. Vytlacil. Structural equations, treatment effects, and econometric policy evaluation 1. *Econometrica*, 73(3):669–738, 2005.

C. F. Manski. Anatomy of the selection problem. *Journal of Human Resources*, 24(3):343–360, 1989.

C. F. Manski. *Partial Identification of Probability Distributions*. Springer-Verlag, New York, 2003.

C. F. Manski and J. V. Pepper. Monotone instrumental variables: with an application to the returns to schooling. *Econometrica*, 68(4):997–1010, 2000.

M. Mogstad, A. Santos, and A. Torgovitsky. Using instrumental variables for inference about policy relevant treatment parameters. *Econometrica*, 86(5): 1589–1619, 2018.

Part III

Widespread Causal Inference Designs

12

Panel Data

Let (Y, X, η, U) be a random vector where Y, η, and U take values in $\mathbf{R}$ and X takes values in $\mathbf{R}^k$. Note that here we are *not* assuming that the first component of X is a constant equal to one. Let $\beta = (\beta_1, \ldots, \beta_k)' \in \mathbf{R}^k$ be such that

$$Y = X'\beta + \eta + U \ ,$$

where we assume both η and U are unobserved. In addition, we want to allow for the possibility that X and η are correlated, so that $E[X\eta] \neq 0$. Given this, combining $\eta + U$ into a single unobservable would require an IV to get an estimator of β, even if we assume $E[XU] = 0$. Today we will see that when we observe the same units (individuals, firms, families, etc) multiple times (across time, regions, etc) we may identify and consistently estimate β without an IV, at least under certain restrictions on η and U.

Suppose that we observe the same unit at two different points in time, and that the unobservable η captures unobserved heterogeneity that is unit specific *and* constant over time. That is, consider the model

$$\begin{aligned} Y_1 &= X_1'\beta + \eta + U_1 \\ Y_2 &= X_2'\beta + \eta + U_2 \ . \end{aligned}$$

Note that we are also assuming that β is a constant parameter that does not change over time. If this is the case, we could simply take first differences, i.e.,

$$\begin{aligned} Y_2 - Y_1 &= (X_2 - X_1)'\beta + U_2 - U_1 \\ \Delta Y &= \Delta X'\beta + \Delta U \ , \end{aligned}$$

and remove the unobserved individual effect η in the process. Notice that

$$E[\Delta X \Delta U] = E[X_2 U_2] + E[X_1 U_1] - E[X_2 U_1] - E[X_1 U_2] \ . \tag{12.1}$$

For the expression above to be equal to zero it is not enough to assume that $E[X_2 U_2] = E[X_1 U_1] = 0$, which would be the standard orthogonality assumption. We also need that $E[X_2 U_1] = E[X_1 U_2] = 0$, i.e., that the covariates in a given time period are uncorrelated with the unobservables in other time periods. This is called strict exogeneity. If this is the case, running least squares of ΔY on ΔX would deliver a consistent estimator of

$$\beta = E[\Delta X \Delta X']^{-1} E[\Delta X \Delta Y] \ , \tag{12.2}$$

provided that $E[\Delta X \Delta X']$ is invertible.

Before we proceed to formalize and extend some of these ideas, there are a few aspects that are worth keeping in mind. First, observing the same units over multiple time periods (the so-called *panel data*) allow us to control for unobserved factors that are constant over time (the η). The trick we just used would not work if η was allowed to *change over time*. Second, the requirement that $E[\Delta X \Delta X']$ is invertible means that we need X to *change over time*, so the trick we just used does not allow us to estimate coefficients of variables that are constant over time. Indeed, such variables are removed by the transformation in the same way η is removed. Finally, strict exogeneity is arguably stronger than simply assuming $E[X_t U_t] = 0$ for all t. Cases where X_2 is a decision variable of an agent in a context where U_1 is known at $t = 2$ may seriously question the validity of $E[X_2 U_1] = 0$. Note that this type of dynamic argument is distinct from omitted variables bias in the sense that it could occur even if we were to argue that $E[X_t U_t] = 0$ or even $E[U_t | X_t] = 0$.

12.1 Fixed Effects

12.1.1 First Differences

Let (Y, X, η, U) be distributed as described above and denote by P the distribution of

$$(Y_{i,1}, \ldots, Y_{i,T}, X_{i,1}, \ldots, X_{i,T}) \ . \tag{12.3}$$

We assume that we have a random sample of size n, so that the observed data is given by $\{(Y_{i,t}, X_{i,t}) : 1 \leq i \leq n, 1 \leq t \leq T\}$. Note that while the sampling process is i.i.d. across i, we are being completely agnostic about the dependence across time for a given unit i. We then consider

$$Y_{i,t} = X_{i,t}'\beta + \eta_i + U_{i,t}, \ i = 1, \ldots, n; \quad t = 1, \ldots, T \ , \tag{12.4}$$

under the assumptions on $X_{i,t}$ and $U_{i,t}$ that we formalize below. Now define

$$\Delta X_{i,t} = X_{i,t} - X_{i,t-1}$$

for $t \geq 2$, and proceed analogously with the other random variables. Note again that $\Delta \eta_i = 0$. Applying this transformation to (12.4), we get

$$\Delta Y_{i,t} = \Delta X_{i,t}'\beta + \Delta U_{i,t}, \ i = 1, \ldots, n \quad t = 2, \ldots, T \ . \tag{12.5}$$

It follows that a regression of $\Delta Y_{i,t}$ on $\Delta X_{i,t}$ provides a consistent estimator of β if the following two assumptions hold,

FD1. $E[U_{i,t} | X_{i,1}, \ldots, X_{i,T}] = 0$ for all $t = 1, \ldots, T$,

FD2. $\sum_{t=2}^{T} E[\Delta X_{i,t} \Delta X_{i,t}']$ is finite and invertible.

FD1 is a sufficient condition for $E[\Delta U_{i,t}|\Delta X_{i,t}] = 0$. FD2 fails if some component of $X_{i,t}$ does not vary over time. The first-difference estimator then takes the form

$$\hat{\beta}_n^{\text{fd}} = \left(\sum_{1\leq i\leq n} \sum_{2\leq t\leq T} \Delta X_{i,t} \Delta X_{i,t}' \right)^{-1} \left(\sum_{1\leq i\leq n} \sum_{2\leq t\leq T} \Delta X_{i,t} \Delta Y_{i,t} \right) . \quad (12.6)$$

Under the assumption that $\text{Var}[U_{i,t}|X_{i,1},\ldots,X_{i,T}]$ is constant (homoskedasticity), together with the assumption of no serial correlation in $U_{i,t}$, it is possible to show that $\hat{\beta}_n^{\text{fd}}$ is not asymptotically efficient and that a different transformation of the data delivers an estimator with a lower asymptotic variance under those assumptions. We will discuss this further after describing this alternative transformation.

12.1.2 Deviations from Means

An alternative transformation to remove the individual effects η_i from (12.4) is the so-called de-meaning technique. In order to define this formally, let

$$\dot{X}_{i,t} = X_{i,t} - \bar{X}_i \quad \text{where} \quad \bar{X}_i = \frac{1}{T} \sum_{1\leq t\leq T} X_{i,t} ,$$

and define $\dot{Y}_{i,t}$ and $\dot{U}_{i,t}$ analogously. Note that $\dot{\eta}_i = 0$ for all $i = 1,\ldots,n$. Applying this transformation to (12.4), we get

$$\dot{Y}_{i,t} = \dot{X}_{i,t}'\beta + \dot{U}_{i,t}, \quad i = 1,\ldots,n \quad t = 1,\ldots,T . \quad (12.7)$$

It follows that a regression of $\dot{Y}_{i,t}$ on $\dot{X}_{i,t}$ provides a consistent estimator of β if the following two assumptions hold,

FE1. $E[U_{i,t}|X_{i,1},\ldots,X_{i,T}] = 0$ for all $t = 1,\ldots,T$,

FE2. $\sum_{t=1}^{T} E[\dot{X}_{i,t}\dot{X}_{i,t}']$ is finite and invertible.

FE1, which is the same strict exogeneity condition in FD1, is a sufficient condition for $E[\dot{U}_{i,t}|\dot{X}_{i,t}] = 0$. As before, FE2 fails if some component of $X_{i,t}$ does not vary over time. The de-meaning estimator (commonly known as the fixed effects estimator, or dummy variable estimator) takes the form

$$\hat{\beta}_n^{\text{fe}} = \left(\sum_{1\leq i\leq n} \sum_{1\leq t\leq T} \dot{X}_{i,t} \dot{X}_{i,t}' \right)^{-1} \left(\sum_{1\leq i\leq n} \sum_{1\leq t\leq T} \dot{X}_{i,t} \dot{Y}_{i,t} \right) . \quad (12.8)$$

Under the assumption that $\text{Var}[U_{i,t}|X_{i,1},\ldots,X_{i,T}]$ is constant (homoskedasticity), together with the assumption of no serial correlation in $U_{i,t}$, it is possible to show that $\hat{\beta}_n^{\text{fe}}$ is asymptotically efficient. The name dummy variable

estimator comes from the fact that the same coefficient on $X_{i,t}$ is obtained by running a regression of $Y_{i,t}$ on $X_{i,t}$ and a full set of unit dummies (with no intercept); the Frisch-Waugh-Lovell theorem leaves exactly the demeaned variables $\dot{Y}_{i,t}$ and $\dot{X}_{i,t}$. We discuss efficiency in the next section.

12.1.3 Asymptotic Properties

Deriving an asymptotic approximation for estimators in panel data models involves two elements that were not present with cross-sectional data. First, the data is i.i.d. across i but may be dependent across time. That is, we may suspect that $X_{i,t}$ and $X_{i,s}$ for $t \neq s$ may not be independent. Second, the data has two indices now: the number of units (denoted by n) and the number of time periods (denoted by T). We will definitely need $nT \to \infty$ to get a useful asymptotic approximation, but we may achieve this by all sorts of different assumptions about how n and/or T grow. The two standard approximations are $n \to \infty$ and T fixed (the so-called short panels) and $n \to \infty$ and $T \to \infty$ (the so-called large panels). Many commonly used panels in applied research include thousands of units (n large) and few time periods (T small) so we will focus on short panels first and discuss large panels later in class.

Under asymptotics where $n \to \infty$ and fixed T, we can show that $\hat{\beta}_n^{\text{fe}}$ and $\hat{\beta}_n^{\text{fd}}$ are asymptotically normal using similar arguments to those we use before, provided we assume

$$(Y_{i,1}, \ldots, Y_{i,T}, X_{i,1}, \ldots, X_{i,T}, U_{i,1}, \ldots, U_{i,T})$$

are i.i.d. across $i = 1, \ldots, n$. Start by writing

$$\sqrt{n}(\hat{\beta}_n^{\text{fe}} - \beta) = \left(\frac{1}{n} \sum_{1 \leq i \leq n} \sum_{1 \leq t \leq T} \dot{X}_{i,t} \dot{X}_{i,t}' \right)^{-1} \left(\frac{1}{\sqrt{n}} \sum_{1 \leq i \leq n} \sum_{1 \leq t \leq T} \dot{X}_{i,t} \dot{U}_{i,t} \right) .$$

In order to make this expression more tractable, we use two tricks. First, note that

$$\sum_{1 \leq t \leq T} \dot{X}_{i,t} \dot{U}_{i,t} = \sum_{1 \leq t \leq T} \dot{X}_{i,t} U_{i,t} - \bar{U}_i \sum_{1 \leq t \leq T} \dot{X}_{i,t} = \sum_{1 \leq t \leq T} \dot{X}_{i,t} U_{i,t} , \qquad (12.9)$$

where the last step follows from $\sum_{1 \leq t \leq T} \dot{X}_{i,t} = 0$. We can therefore replace $\dot{U}_{i,t}$ with $U_{i,t}$. Second, let $\dot{X}_i$ be the $T \times k$ matrix whose t-th row is $\dot{X}_{i,t}'$, and define U_i as the T-dimensional vector of stacked observations for unit i. Using this notation, we can write

$$\dot{X}_i' \dot{X}_i = \sum_{1 \leq t \leq T} \dot{X}_{i,t} \dot{X}_{i,t}' \quad \text{and} \quad \dot{X}_i' U_i = \sum_{1 \leq t \leq T} \dot{X}_{i,t} U_{i,t} . \qquad (12.10)$$

Combining (12.9) and (12.10), we obtain

$$\sqrt{n}(\hat{\beta}_n^{\text{fe}} - \beta) = \left(\frac{1}{n} \sum_{1 \leq i \leq n} \dot{X}_i' \dot{X}_i \right)^{-1} \left(\frac{1}{\sqrt{n}} \sum_{1 \leq i \leq n} \dot{X}_i' U_i \right) .$$

By the law of large numbers and FE2,

$$\frac{1}{n}\sum_{1\leq i\leq n}\dot{X}_i'\dot{X}_i \xrightarrow{P} \Sigma_{\dot{X}} \equiv E[\dot{X}_i'\dot{X}_i] = \sum_{1\leq t\leq T} E[\dot{X}_{i,t}\dot{X}_{i,t}'] \ .$$

In addition, by the central limit theorem and FE1,

$$\frac{1}{\sqrt{n}}\sum_{1\leq i\leq n}\dot{X}_i'U_i \xrightarrow{d} N(0,\Omega), \quad \text{where} \quad \Omega = \mathrm{Var}[\dot{X}_i'U_i] = E[\dot{X}_i'U_iU_i'\dot{X}_i] \ .$$

Combining these results with the CMT we get

$$\sqrt{n}(\hat{\beta}_n^{\mathrm{fe}} - \beta) \xrightarrow{d} N(0, \mathbb{V}^{\mathrm{fe}}) \tag{12.11}$$

where

$$\mathbb{V}^{\mathrm{fe}} = \Sigma_{\dot{X}}^{-1}\Omega\Sigma_{\dot{X}}^{-1} \ . \tag{12.12}$$

Historically, researchers often assumed that $U_{i,t}$ was serially uncorrelated with variance independent of $X_{i,t}$ (i.e. homoskedastic). The default standard errors in Stata are still based on these assumptions. However, these assumptions are difficult to justify for most economic data, which is often strongly autocorrelated and heteroskedastic. One faces basically the same trade-off as with heteroskedasticity in the cross-sectional case. The most common strategy is to use the fully robust consistent estimator of the asymptotic variance,

$$\hat{\mathbb{V}}^{\mathrm{fe}} = \left(\frac{1}{n}\sum_{1\leq i\leq n}\dot{X}_i'\dot{X}_i\right)^{-1}\left(\frac{1}{n}\sum_{1\leq i\leq n}\dot{X}_i'\hat{U}_i\hat{U}_i'\dot{X}_i\right)\left(\frac{1}{n}\sum_{1\leq i\leq n}\dot{X}_i'\dot{X}_i\right)^{-1} \ ,$$

where $\hat{U}_i = \dot{Y}_i - \dot{X}_i\hat{\beta}_n^{\mathrm{fe}}$. This is what Stata computes when one uses the `cluster(unit)` option to `xtreg` where unit is the variable that indexes i. This estimator is an appealing generalization of White's (1980) heteroskedasticity consistent covariance matrix estimator that allows for arbitrary inter-temporal correlation patterns and heteroskedasticity across individuals. As we will see later in the cluster covariance material, this estimator is generally known as a cluster covariance estimator (CCE) and is consistent as $n \to \infty$, i.e., $\hat{\mathbb{V}}^{\mathrm{fe}} \xrightarrow{P} \mathbb{V}^{\mathrm{fe}}$.

12.1.3.0.1 A comment on efficiency.

Traditional arguments in favor of the fixed effects (or within-group) estimator $\hat{\beta}_n^{\mathrm{fe}}$ over the first-difference estimator $\hat{\beta}_n^{\mathrm{fd}}$ rely on the fact that under homoskedasticity and no-serial correlation of $U_{i,t}$, $\hat{\beta}_n^{\mathrm{fe}}$ has a lower asymptotic variance than $\hat{\beta}_n^{\mathrm{fd}}$. In that case, the covariance matrix of U_i is proportional to the identity, so the within transformation is the GLS transformation among transformations that remove η_i. By contrast, first differencing creates

a moving-average error even when the original $U_{i,t}$ is serially uncorrelated. Intuitively, taking first differences introduces correlation in $\Delta U_{i,t}$ as

$$\begin{aligned} E[\Delta U_{i,t}\Delta U_{i,t-1}] &= E[U_{i,t}U_{i,t-1} - U_{i,t-1}U_{i,t-1} - U_{i,t}U_{i,t-2} + U_{i,t-1}U_{i,t-2}] \\ &= -\operatorname{Var}(U_{i,t-1}) \ . \end{aligned}$$

However, in the other extreme where $U_{i,t}$ follows a random walk, i.e., $U_{i,t} = U_{i,t-1} + V_{i,t}$ for some i.i.d. sequence $V_{i,t}$, then $\Delta U_{i,t} = V_{i,t}$. These results, at the end of the day, rely on homoskedasticity and so it is advised to simply use a robust standard error as above and forget about efficiency considerations. Note that when $T = 2$, these two estimators are numerically the same. In addition, first differences are used in dynamic panels and difference in differences, as we will discuss later.

Remark 12.1 Panel data traditionally deals with units over time. However, we can think about other cases where the data has a two-dimensional index and where we believe that one of the indices may exhibit within group dependence. For example, it could be that we observe "employees" within "firms", or "students" within "schools", or "families" in metropolitan statistical areas (MSA), etc. Cases like these are similar but not identical to panel data. To start, units are not "repeated" in the sense that each unit is potentially observed only once in the sample. In addition, these are cases where "T" is usually large and "n" is small. For example, we typically observe many students (which may be dependent within a school) and few schools. We will study these cases later in the class. ■

12.2 Random Effects

Fixed effects approaches are attractive to economists because they provide a way of addressing omitted variables bias and related forms of endogeneity, as long as the omitted factors are time constant. An alternative way to exploit the time dimension of the panel is to model the evolution of the unobservable term over time within a unit, and use this model to increase efficiency relative to ordinary pooled linear regressions. This is known as a random effects approach. Random effects are not as widely used as fixed effects in economics because they focus on efficiency rather than bias and robustness. Nevertheless, random effects approaches are occasionally used and also have some interesting connections to fixed effects and other types of panel data models.

The standard random effects model adds the following assumption to (12.4),

RE1. $E[\eta_i | X_{i,1}, \ldots, X_{i,T}] = 0$.

Hence all of the unobservable time-invariant factors that were being controlled for in the fixed effects approach are now assumed to be mean independent (ergo, uncorrelated) with the explanatory variables at all time periods. The strict exogeneity condition of the fixed effects approach (i.e. FE1) is still maintained, so that the aggregate error term $V_{i,t} = \eta_i + U_{i,t}$ now satisfies $E[V_{i,t}|X_{i,1}, \ldots, X_{i,T}] = 0$ for all $t = 1, \ldots, T$. The idea behind the random effects approach is to exploit the serial correlation in $V_{i,t}$ that is generated by having a common η_i component in each time period. Specifically, the baseline approach maintains the following.

RE2. (i) $\text{Var}[U_{i,t}|X_{i,1}, \ldots, X_{i,T}] = \sigma_U^2$, (ii) $\text{Var}[\eta_i|X_{i,1}, \ldots, X_{i,T}] = \sigma_\eta^2$, (iii) $E[U_{i,t}U_{i,s}|X_{i,1}, \ldots, X_{i,T}] = 0$ for all $t \neq s$, (iv) $E[U_{i,t}\eta_i|X_{i,1}, \ldots, X_{i,T}] = 0$ for all $t = 1, \ldots, T$.

Under these assumptions,

$$\text{Var}[V_{i,t}|X_{i,1}, \ldots, X_{i,T}] = E[\eta_i^2 + U_{i,t}^2 + 2\eta_i U_{i,t}|X_{i,1}, \ldots, X_{i,T}] = \sigma_\eta^2 + \sigma_U^2 \ ,$$

and

$$E[V_{i,t}V_{i,s}|X_{i,1}, \ldots, X_{i,T}] = E[\eta_i^2 + U_{i,t}U_{i,s} + \eta_i U_{i,t} + \eta_i U_{i,s}|X_{i,1}, \ldots, X_{i,T}] = \sigma_\eta^2 \ .$$

Combining these results and stacking the observations for unit i, we get that

$$E[V_i V_i'|X_i] = \Omega = \sigma_U^2 \mathbb{I}_T + \sigma_\eta^2 \iota_T \iota_T' \ , \tag{12.13}$$

where $\mathbb{I}_T$ is the $T \times T$ identity matrix and ι_T is a T-dimensional vector of ones. Under these assumptions, the estimator with the lowest asymptotic variance is

$$\hat{\beta}_n^{\text{re}} = \left(\sum_{1 \leq i \leq n} X_i' \Omega^{-1} X_i \right)^{-1} \left(\sum_{1 \leq i \leq n} X_i' \Omega^{-1} Y_i \right) \ , \tag{12.14}$$

where X_i is the $T \times k$ matrix of stacked observations for unit i, and Y_i is the corresponding T-dimensional vector. Note this is just a generalized least squares (GLS) estimator of β. This GLS estimator is, nevertheless, unfeasible, since Ω depends on the unknown parameters σ_U^2 and σ_η^2. However, these two can be easily estimated to form $\hat{\Omega}$ and deliver a feasible GLS estimator of β.

A few aspects are worth discussing. First, the efficiency gains hold under the additional structure imposed by RE1 and RE2. In particular, we are possibly gaining efficiency in a context where the unobserved heterogeneity η_i is assumed to be mean independent of X_i. In other words, unobserved time-invariant factors must be uncorrelated with observed covariates. This was precisely what made the fixed effects approach attractive in the first place. Second, the efficiency gains hold under the homoskedasticity and independence assumptions in RE2 and do not hold more generally. These are undoubtedly strong assumptions. Third, unlike the fixed effects estimator, the random effects approach allows to estimate regression coefficients associated with time-invariant covariates (that is, some of the $X_{i,t}$ may be constant across time,

such as gender of the individual). So if the analysis is primarily concerned with the effect of a time-invariant regressor and panel data is available, it makes sense to consider some sort of random effects type of approach. Relatedly, correlated random effects approaches, such as the Mundlak or Chamberlain devices, start from the random effects framework but add controls for the time averages or histories of the covariates to relax RE1. Fourth, under RE1 and RE2 β is identified in a single cross-section. The parameters that require panel data for identification in this model are the variances of the components of the error σ_η^2 and σ_U^2, which are needed for the GLS approach. Finally, note that the terminology "fixed effects" and "random effects" is arguably confusing as η_i is random in both approaches.

A last word of caution should be made about the use of Hausman specification tests. These are tests that compare $\hat{\beta}_n^{\text{fe}}$ with $\hat{\beta}_n^{\text{re}}$ in order to test the validity of RE1 (assuming RE2 holds). Under the null hypothesis that RE1 holds, both estimators are consistent but $\hat{\beta}_n^{\text{re}}$ is efficient. Under the alternative hypothesis, $\hat{\beta}_n^{\text{fe}}$ is consistent while $\hat{\beta}_n^{\text{re}}$ is not. Now, suppose we were to define a new estimator $\hat{\beta}_n^*$ as follows

$$\hat{\beta}_n^* = \hat{\beta}_n^{\text{fe}} I\{\text{Hausman test rejects}\} + \hat{\beta}_n^{\text{re}} I\{\text{Hausman test accepts}\} \ . \quad (12.15)$$

The problem with this new estimator is that its finite sample distribution looks very different from the usual normal approximations. This is generally the case when there is pre-testing, understood as a situation where we conduct a test in a first step, and then depending on the outcome of this test, we do A or B in a second step. Near the boundary between the null and the alternative, the decision to use FE or RE is itself random, so the resulting distribution can look like a mixture rather than a single normal approximation. A formal analysis of these *uniformity* issues is covered in 481 and is beyond the scope of this class.

12.3 Dynamic Models

One benefit of panel data is that it allows us to analyze economic relationships that are inherently dynamic. Specifically, we may be interested in the effect of lagged outcomes on future outcomes. Let $\{Y_{i,t} : 1 \leq i \leq n, 0 \leq t \leq T\}$ be a sequence of random variables and consider the model

$$Y_{i,t} = \rho Y_{i,t-1} + \eta_i + U_{i,t}, \quad i = 1, \ldots, n \quad t = 1, \ldots, T \ , \quad (12.16)$$

where η_i and $U_{i,t}$ are the same as before but now $Y_{i,t-1}$ is allowed to have a direct effect on $Y_{i,t}$, a feature sometimes referred to as state dependence. We assume that $|\rho| < 1$. As is common in dynamic panel data (and time series) contexts, we will assume that the model is dynamically complete in the sense

that all appropriate lags of $Y_{i,t}$ have been removed from the time-varying error $U_{i,t}$, i.e.,

$$E[U_{i,t}|Y_{i,t-1}, Y_{i,t-2}, \dots] = 0 \text{ for all } t = 1, \dots, T. \tag{12.17}$$

Consider now taking first differences to (12.16) to obtain,

$$\Delta Y_{i,t} = \rho \Delta Y_{i,t-1} + \Delta U_{i,t}, \quad i = 1, \dots, n \quad t = 2, \dots, T \ ,$$

where, as before, $\Delta Y_{i,t} = Y_{i,t} - Y_{i,t-1}$ and similarly for $U_{i,t}$. Here we assume the data also include the initial value $Y_{i,0}$, which we treat as observed. In general we will have $\text{Cov}(\Delta Y_{i,t-1}, \Delta U_{i,t}) \neq 0$ since (12.16) implies

$$\text{Cov}(Y_{i,t-1}, U_{i,t-1}) \neq 0 \ . \tag{12.18}$$

A similar conclusion would arise if we tried to use the de-meaning transformation. This inherent endogeneity is a generic feature of models that have both state dependence and time-invariant heterogeneity. In order to get rid of the fixed effects we have to compare outcomes over time, but if past outcomes have effects on future outcomes then differenced error terms will still be correlated with the differenced lagged outcomes that appear as regressors.

The most commonly proposed solution to this problem is to use other lagged outcomes as instruments. Given (12.17), we know that $Y_{i,t-2}$ is uncorrelated with both $U_{i,t}$ and $U_{i,t-1}$, hence

$$\text{Cov}(Y_{i,t-2}, \Delta U_{i,t}) = 0 \ .$$

At the same time, we also know that

$$\begin{aligned} \text{Cov}(Y_{i,t-2}, \Delta Y_{i,t-1}) &= \text{Cov}(Y_{i,t-2}, Y_{i,t-1}) - \text{Cov}(Y_{i,t-2}, Y_{i,t-2}) \\ &= \text{Cov}(Y_{i,t-2}, \rho Y_{i,t-2} + \eta_i + U_{i,t-1}) - \text{Cov}(Y_{i,t-2}, Y_{i,t-2}) \\ &= -(1-\rho)\,\text{Var}[Y_{i,t-2}] + \text{Cov}(Y_{i,t-2}, \eta_i) \ , \end{aligned}$$

The covariance above is generically nonzero whenever $|\rho| < 1$ and $\text{Var}[Y_{i,t-2}] > 0$, which makes $Y_{i,t-2}$ a relevant instrument for $\Delta Y_{i,t-1}$ while dynamic completeness makes it exogenous for $\Delta U_{i,t}$. An actual expression for this last covariance can be obtained under additional assumptions. For example, under the assumption that the initial condition, $Y_{i,0}$, is independent of η_i (and $\eta_i \perp U_{i,t}$), then

$$\text{Cov}(Y_{i,t-2}, \eta_i) = \sigma_\eta^2 \sum_{j=0}^{t-3} \rho^j \ .$$

This strategy requires $T \geq 3$, since otherwise we would not have data on $Y_{i,t-2}$. For larger T we could include additional lags such as $Y_{i,t-3}, Y_{i,t-4}$, etc. Following such an approach delivers $(T-2)(T-1)/2$ linear moment restrictions of the form

$$E[Y_{i,t-k}(\Delta Y_{i,t} - \rho \Delta Y_{i,t-1})] = 0, \quad t = 3, \dots, T, \quad k = 2, \dots, t-1 \ . \tag{12.19}$$

The predictive power of these lags for $\Delta Y_{i,t-1}$ is likely to get progressively weaker as the lag distance gets larger. Weak instrument problems may arise as a consequence. If $T \geq 4$ then one could consider using the differenced term $\Delta Y_{i,t-2}$ (instead or in addition to the level $Y_{i,t-2}$) as an instrument for $\Delta Y_{i,t-1}$. In practice, these moment restrictions are typically combined in a GMM estimator that chooses ρ to make sample analogs of the moments above as close to zero as possible. In the literature, these approaches are frequently referred to as Arellano-Bond or Anderson-Hsiao estimators; see Arellano [2003]. The key point is that directly applying within-group OLS to (12.16) is not a good short-panel solution: it suffers from Nickell bias of order $1/T$.

12.4 Acknowledgement

The material today borrows from several useful sources, including lecture notes kindly shared by Alex Torgovitsky. I want to particularly thank Alex for sharing his notes with me.

12.5 Problems

Problem 12.1 *Show that when $T = 2$, the FE and FD estimators of β are numerically the same.*

Problem 12.2 *Show that a projection of a random variable $W_{i,t}$ on fixed and time effects is equal to $\bar{W}_i + \bar{W}_t - \bar{W}$, where $\bar{W}_i$ is an average over time, $\bar{W}_t$ is an average over individuals, and $\bar{W}$ is a full sample average.*

Problem 12.3 *Consider the panel data model in (12.4). Define a full set of unit dummies $\{D_{i,t}^{(j)} = I\{i = j\} : j = 1, \ldots, n\}$ and let $\hat{\beta}_n^{\text{dum}}$ denote the coefficient on $X_{i,t}$ in the OLS regression of $Y_{i,t}$ on $X_{i,t}$ together with the n unit dummies (with no overall intercept). Show that*

$$\hat{\beta}_n^{\text{dum}} = \hat{\beta}_n^{\text{fe}} \ ,$$

where $\hat{\beta}_n^{\text{fe}}$ is the within-group estimator defined in the "Deviations from Means" subsection. Hint: apply the Frisch-Waugh-Lovell theorem to the dummy-variable regression. The residual from regressing $X_{i,t}$ on the unit dummies is exactly $\dot{X}_{i,t}$.

Problem 12.4 *Consider the dynamic panel model*

$$Y_{i,t} = \rho Y_{i,t-1} + \eta_i + U_{i,t}, \quad i = 1, \dots, n, \quad t = 1, \dots, T \ ,$$

with $|\rho| < 1$*, dynamic completeness as in* (12.17)*,* η_i *independent of* $\{U_{i,t}\}_{t \geq 1}$ *and the initial condition* $Y_{i,0}$*, and stationarity in the sense that* $\mathrm{Var}[Y_{i,t}]$ *does not depend on* t*. Let* $\hat{\rho}_n^{\mathrm{fe}}$ *be the within-group estimator obtained by applying* $(\dot{\cdot})$ *to both sides of the model and regressing* $\dot{Y}_{i,t}$ *on* $\dot{Y}_{i,t-1}$*.*

(a) Explain in words why $\hat{\rho}_n^{\mathrm{fe}}$ *is biased and inconsistent for* ρ *when* T *is fixed and* $n \to \infty$*. (Identify the source of the correlation between* $\dot{Y}_{i,t-1}$ *and* $\dot{U}_{i,t}$*.)*

(b) Show that, under fixed T *and* $n \to \infty$*,*

$$\operatorname{plim} \hat{\rho}_n^{\mathrm{fe}} - \rho = -\frac{1+\rho}{T-1} \, \frac{1 - \frac{1}{T} \, \frac{1-\rho^T}{1-\rho}}{1 - \frac{2\rho}{(1-\rho)(T-1)} \left(1 - \frac{1}{T} \, \frac{1-\rho^T}{1-\rho}\right)} \ .$$

This is the Nickell bias *(Nickell, 1981). Conclude that within-group OLS is not a sensible estimator of* ρ *in short panels.*

(c) Evaluate the bias for $\rho = 0.5$ *and* $T = 3$*,* $T = 10$*, and* $T = 50$*. Comment.*

Bibliography

M. Arellano. *Panel Data Econometrics*. Oxford University Press, 2003.

J. M. Wooldridge. *Econometric analysis of cross section and panel data*. MIT press, 2010.

13

Difference in Differences

Today we focus again on the problem of evaluating the impact of a program or treatment on a population outcome Y. Instead of relying on data from a randomized experiment, assumptions like selection on observables, or the availability of an instrument, we will focus attention on natural experiments with panel data. Natural experiments used in the social sciences are often policy changes. For instance, a US state increases its minimum wage while the neighboring state does not, thus giving researchers a treatment group facing a high minimum wage, and a control group facing a lower minimum wage. Natural experiments often affect an entire state, region, or province, so findings from studies leveraging natural experiments typically apply to large and unselected populations, unlike findings from randomized experiments. However, in natural experiments, assignment to the treatment is not randomized by a researcher, it is decided by a policy maker. Since policy makers do not randomly choose where to implement a policy change, treated and control locations may not be comparable and the identification of causal parameters becomes less straightforward.

Our goal in this chapter is to discuss the basic idea behind the difference in differences (DiD) approach, a popular method to estimate causal effects in the context of natural experiments. The difference-in-differences estimator compares the change in outcomes for treated groups to the change in outcomes for control groups. We start with a simple illustration in a model with two groups and two time periods, to then move to the more general case.

13.1 Two Groups and Two Periods

The simplest setup to describe the DiD approach is one where outcomes are observed for two groups for two time periods. One of the groups is exposed to a treatment in the second period but not in the first period. The second group is not exposed to the treatment during either period. To be specific, let

$$\{(Y_{g,t}, D_{g,t}) : g \in \{s, n\} \text{ and } t \in \{1, 2\}\} \tag{13.1}$$

be the observed data, where $Y_{g,t} \in \mathbf{R}$ denotes the outcome of interest for group g at time t and $D_{g,t} \in \{0, 1\}$ the treatment status of group g at time

t. A group could be a location, a group of firms, a family, etc. In this simple example, location s switches from untreated to treated from period 1 to 2, and location n is untreated at both dates. This means $D_{g,t} = 1$ if and only if $g = s$ and $t = 2$. Let also $Y_{g,t}(1)$ and $Y_{g,t}(0)$ denote the counterfactual outcomes of group g at t with and without treatment, respectively.

One possible parameter that we may be interested in identifying is the average treatment effect on the treated. In this simple example there is only one such group that is treated in the second time period, so this simplifies to

$$\theta_{\text{att}} = E[Y_{g,t}(1) - Y_{g,t}(0) \mid D_{g,t} = 1] = E[Y_{s,2}(1) - Y_{s,2}(0)] \ . \tag{13.2}$$

Of course we may be interested in other types of causal parameters, including the usual *average treatment effect* (ATE). However, in the framework we study today identifying the ATE would require additional assumptions on the exogeneity of $D_{g,t}$ that are typically difficult to defend in the type of natural experiments where DiD tools are appropriate. We discuss this further in the next section. For the moment, consider the following well-known example as an illustration of the 2×2 simple case of this section.

Example 13.1 On April 1, 1992, New Jersey raised the state minimum wage from \$4.25 to \$5.05. Card and Krueger [1994] collected data on employment at fast food restaurants in New Jersey in February 1992 ($t = 1$) and again in November 1992 ($t = 2$) to study the effect of increasing the minimum wage on employment. They also collected data from the same type of restaurants in eastern Pennsylvania, just across the river. The minimum wage in Pennsylvania stayed at \$4.25 throughout this period. In our notation, New Jersey would be the group s that "switched", $Y_{g,t}$ would be the employment rate in location g at time t, and $D_{g,t}$ denotes an increase in the minimum wage (the treatment) in group g at time t. ■

The identification strategy of DiD relies on a parallel-trends assumption: in the absence of the treatment, both locations would have experienced the same average outcome evolution. Mathematically,

$$E[Y_{s,2}(0) - Y_{s,1}(0)] = E[Y_{n,2}(0) - Y_{n,1}(0)] \ , \tag{13.3}$$

i.e., both groups have "common trends" in the absence of a treatment. As we will properly formalize later in class, the parallel-trends assumption implies that

$$E[Y_{g,t}(0)] = \eta_g + \gamma_t \ , \tag{13.4}$$

where η_g and γ_t are (non-random) group and time effects. This additive structure for non-treated potential outcomes implies that $E[Y_{g,2}(0) - Y_{g,1}(0)] = \gamma_2 - \gamma_1 \equiv \gamma$, which is constant across groups. Note that this assumption, together with (13.2), implies that

$$E[Y_{s,2}(1)] = \theta_{\text{att}} + \eta_s + \gamma_2 \ . \tag{13.5}$$

In the context of the previous example, this assumption says that in the absence of a minimum wage change, the expected employment is determined by the sum of a time-invariant state effect and a year effect that is common across states. Before we discuss the identifying power of this structure, we discuss two natural (but unsuccessful) approaches that may come to mind.

13.1.1 Pre and post comparison

A natural approach to identify θ_{att} in (13.2) would be to compare $Y_{s,2}$ and $Y_{s,1}$; that is, using outcomes before and after the policy change for the treated group alone. This approach delivers,

$$E[\Delta Y_{s,2}] = E[Y_{s,2}(1) - Y_{s,1}(0)] = \theta_{\text{att}} + \gamma \ ,$$

where $\Delta Y_{s,2} = Y_{s,2} - Y_{s,1}$ and $\gamma = \gamma_2 - \gamma_1$. Clearly, this approach does not identify θ_{att} in the presence of time trends, i.e., $\gamma \neq 0$. In the context of Example 13.1, the employment rate in New Jersey may have been going up (or down) in the absence of a policy change (the treatment), and so before and after comparisons confound the time trend as being part of the treatment effect. Unless one is willing to assume $\gamma = 0$, this approach does not identify the ATT.

13.1.2 Treatment and control comparison

A second natural approach to identify θ_{att} in (13.2) would be to compare $Y_{s,2}$ and $Y_{n,2}$; that is, using outcomes from both groups in the second time period. This approach delivers,

$$E[Y_{s,2} - Y_{n,2}] = E[Y_{s,2}(1) - Y_{n,2}(0)] = \theta_{\text{att}} + \eta \ ,$$

where $\eta = \eta_s - \eta_n$. Clearly, this approach does not identify θ_{att} in the presence of persistent group differences, i.e., $\eta \neq 0$. In the context of Example 13.1, the employment rate in New Jersey and Pennsylvania may be idiosyncratically different in the absence of a policy change and so comparing these two states confounds these permanent differences as being part of the treatment effect. Unless one is willing to assume $\eta = 0$, this approach does not identify the ATT.

13.1.3 Taking both differences

The DiD approach exploits the common trends assumption in (13.3) to identify θ_{att}. The idea is to consider a second "difference" to remove γ (the time trend) from the difference that arises from comparing pre and post outcomes. In other words, the idea is to take the "difference" of the "differences", $\Delta Y_{s,2}$ and

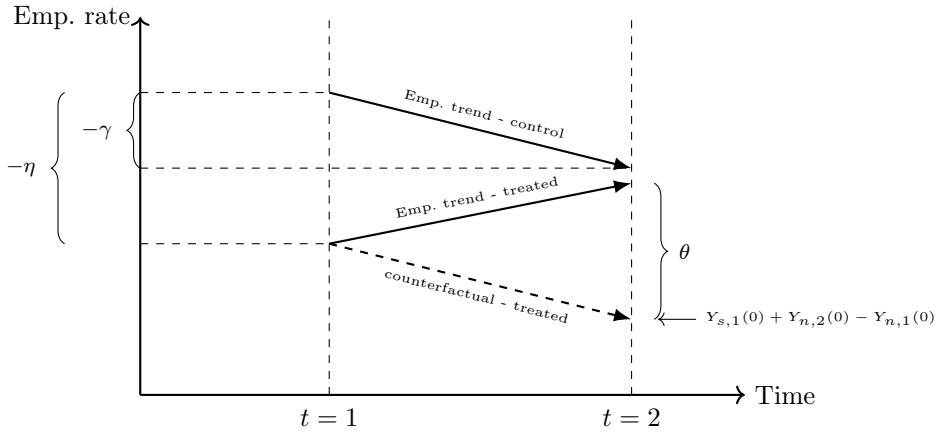


FIGURE 13.1: Causal effects in the DiD model. The brace labeled θ is the treatment effect for the treated group in period 2, while the arrow indicates the counterfactual untreated outcome constructed from the control group's trend.

$\Delta Y_{n,2}$, to obtain

$$\begin{aligned} E[\Delta Y_{s,2} - \Delta Y_{n,2}] &= E[Y_{s,2}(1) - Y_{s,1}(0)] - E[Y_{n,2}(0) - Y_{n,1}(0)] \\ &= \theta_{\text{att}} + \gamma - \gamma = \theta_{\text{att}} \ . \end{aligned}$$

Thus, the approach identifies the treatment effect by taking the differences between pre-versus-post comparisons in the two groups, and exploiting the fact that the time trend γ is "common" in the two groups.

Note that an alternative interpretation to the same idea is to compare $(Y_{s,2} - Y_{n,2})$ and $(Y_{s,1} - Y_{n,1})$, that is, the treatment and control comparison before and after the policy change. This is because

$$\begin{aligned} E[(Y_{s,2} - Y_{n,2}) - (Y_{s,1} - Y_{n,1})] &= E[Y_{s,2}(1) - Y_{n,2}(0)] - E[Y_{s,1}(0) - Y_{n,1}(0)] \\ &= \theta_{\text{att}} + \eta - \eta = \theta_{\text{att}} \ . \end{aligned}$$

Using this representation, the difference for the pre-period is used to remove the persistent group difference η, a strategy that again works under the common trends assumption in (13.3).

A final interpretation of the same idea is that the DiD approach constructs a counterfactual potential outcome $Y_{s,2}(0)$ (which is unobserved) by combining $Y_{s,1}(0)$, $Y_{n,2}(0)$, and $Y_{n,1}(0)$, which are all observed. The "constructed" potential outcome is simply $\tilde{Y}_{s,2}(0) = Y_{s,1}(0) + Y_{n,2}(0) - Y_{n,1}(0)$, so that

$$\begin{aligned} E[\tilde{Y}_{s,2}(0)] &= E[Y_{s,1}(0) + Y_{n,2}(0) - Y_{n,1}(0)] \\ &= \eta_s + \gamma_1 + \eta_n + \gamma_2 - (\eta_n + \gamma_1) \\ &= \eta_s + \gamma_2, \end{aligned}$$

Computing $E[Y_{s,2} - \tilde{Y}_{s,2}(0)] = \theta_{\text{att}}$ therefore delivers a valid identification strategy. Figure 13.1 illustrates this idea.

13.2 Standard Framework in DiD Models

Consider the case where we observe a group-level panel dataset with groups g taking values in G, the set collecting all groups, and periods t taking values in T, the set of all time periods. For simplicity, we assume that the group-level panel dataset is balanced: the outcome and treatment of each group is observed at every period, which explains why T is not indexed by g. We denote the number of groups by $|G|$ and the number of time periods by $|T|$. Typically, groups are locations, like states, counties, or municipalities, but a group could also be a subset of individuals defined by time-invariant characteristics or be a single individual or firm.

We focus on the case where the treatment D is assigned at the (g,t) level, as is the case when the treatment is a county-level law or regulation, like the minimum wage. When the treatment varies within (g,t) cells, the design is called *fuzzy*. This may arise when groups are geographical entities and individuals or firms within the same cell may not all have the same treatment; in such cases, $D_{g,t}$ typically denotes the average treatment in cell (g,t). Here we do not consider the fuzzy case. We also focus on the case where the treatment is **binary**, $D_{g,t} \in \{0,1\}$. Several of the results we discuss below apply with modest modifications to the case where $D_{g,t}$ takes multiple values, including the case where the treatment is continuous. But these cases are beyond the scope of this class.

In terms of notation, with some abuse of notation we let $D_g \equiv (D_{g,t} : t \in T)$ be a $1 \times |T|$ vector stacking the treatments of group g from period 1 to $|T|$, and let $D^{(n)} = (D_g : g \in G)$ be a vector stacking the treatments of all groups at every period. $D^{(n)}$ is referred to as the design of a study. In addition,

$$\text{for any } (d_1, \ldots, d_{|T|}) \in \{0,1\}^{|T|} \text{ we let } Y_{g,t}(d_1, \ldots, d_{|T|}) \tag{13.6}$$

denote the *potential outcome* of group g at t and let $Y_{g,t} = Y_{g,t}(D_g)$ denote the *observed* outcome of g at t. This *dynamic* potential outcome framework explicitly allows groups' outcome at time t to depend on their past and future treatments. For instance, if the treatment is a policy change, the outcome of a group may depend on the group's treatment history. Soon we will introduce assumptions that will simplify this notation.

Remark 13.1 In most of the literature on DiD, the study design $D^{(n)}$ is implicitly conditioned upon. This is in line with the focus on non-randomized natural experiments, where the researcher does not control the assignment to the treatment, and has to take it as a given. We follow this convention here

and make the assumptions below conditional on the design. Concretely, whenever there is an $E[X]$ below, it should actually be understood as $E[X|D^{(n)}]$. Leaving this conditioning implicit greatly alleviates the notational burden. Conditional on the design, groups' potential outcomes are the only source of randomness left. ■

The DiD approach relies on three assumptions: two that simplify the dynamic potential outcome notation to $Y_{g,t}(d_t)$, and one parallel-trends condition. We present the parallel-trends condition in two versions, the more restrictive of which we impose throughout most of this chapter.

Assumption 13.1 (No anticipation) *For all* $(d_1, \ldots, d_{|T|})$ *and* $g \in G$,

$$Y_{g,t}(d_1, \ldots, d_{|T|}) = Y_{g,t}(d_1, \ldots, d_t) \ .$$

Assumption 13.2 (No dynamics) *For all* $(d_1, \ldots, d_t)$ *and* $g \in G$,

$$Y_{g,t}(d_1, \ldots, d_t) = Y_{g,t}(d_t) \ .$$

Assumption 13.1 requires that a group's potential outcomes do not depend on their future treatments. This assumption may be violated when, for example, a policy change is announced with anticipation and agents behave at t accounting for the future policy change. Assumption 13.2 requires that a group's current outcomes do not depend on its past treatments. This assumption is violated, for example, when the length of exposure to treatment affects the outcome. Under these two assumptions, and with a binary treatment, each cell (g, t) has two potential outcomes: $Y_{g,t}(0)$ if g is untreated at t, and $Y_{g,t}(1)$ if g is treated at t. Then, we are back to the standard Neyman-Rubin model of potential outcomes we have used before.

Our final assumption is the parallel-trends assumption. There are two variations of this assumption we will use. The first one is the more traditional way, which holds when Assumptions 13.1-13.2 are maintained and so potential outcomes take the form $Y_{g,t}(d_t)$.

Assumption 13.3 (Parallel-trends) *For all* $t \geq 2$, $E[Y_{g,t}(0) - Y_{g,t-1}(0)]$ *does not vary across* $g \in G$.

Assumption 13.3 captures the most standard way to describe parallel trends, in a framework where potential outcomes only depend on the treatments that happen in the same period and there is no dependence on past treatments. Later on in class, when we move to event studies, we invoke a more general version of parallel trends that becomes necessary when Assumption 13.2 is not necessarily imposed. In this case, and if we let 0_k denote a k-dimensional vector of zeros, parallel trends can be written as follows.

Assumption 13.4 (General parallel-trends) *For all* $t \geq 2$, $E[Y_{g,t}(0_t) - Y_{g,t-1}(0_{t-1})]$ *does not vary across* $g \in G$.

Again, under Assumptions 13.1-13.2, Assumption 13.4 collapses to 13.3.

Assumptions 13.1-13.3 imply an additively separable structure for the mean potential outcomes for the untreated in terms of a group effect η_g and a time effect γ_t, i.e.,

$$E[Y_{g,t}(0)] = \eta_g + \gamma_t \ . \tag{13.7}$$

The final assumption that becomes particularly relevant when we discuss inference in DiD models is that the potential outcomes of different groups are independent (though not necessarily i.i.d.). That is, we assume that the vectors $\{Y_{g,t}(d_t) : t \in T\}$ are independent across $g \in G$.

13.3 The Two Way Fixed Effects Estimator

In the simple 2×2 model we previously discussed, the DiD contrast,

$$E[\Delta Y_{s,2} - \Delta Y_{n,2}] \ ,$$

can be equivalently obtained as the coefficient in a two-way fixed effects (TWFE) linear regression of $Y_{g,t}$, on group fixed effects, period fixed effects, and the treatment $D_{g,t}$ of group g at period t. This is exactly the within-group logic from the panel data chapter, now applied to a binary treatment variable and with time effects included. Motivated by this fact, researchers have also estimated TWFE regressions in complicated designs with many locations and periods, variation in treatment timing, treatments switching on and off, and/or non-binary treatments. All in all, TWFE regressions are widely popular in empirical work.

Let the observed data be given by $\{(Y_{g,t}, D_{g,t}) : g \in G, t \in T\}$. The TWFE regression consists of the following linear projection,

$$Y_{g,t} = \tilde{\eta}_g + \tilde{\gamma}_t + \beta^{\text{fe}} D_{g,t} + U_{g,t} \ , \tag{13.8}$$

where $\tilde{\eta}_g$ is a group fixed effect, $\tilde{\gamma}_t$ is a time fixed effect, β^{fe} is the coefficient of interest, and $U_{g,t}$ is a projection error. We use tildes to emphasize that these are projection coefficients, not necessarily the structural group and time effects in (13.4). It is important to understand that these regressions are not usually interpreted as a linear model for $Y_{g,t}$ but rather viewed as a mechanical way to compute an estimand β^{fe} that, hopefully, admits an interesting causal interpretation. In particular, we know from the previous section that this is indeed the case in the 2×2 model where β^{fe} equals the ATT.

Despite its massive popularity, it turns out that whether the TWFE estimand β^{fe} equals the ATT (or related parameters) depends delicately on the design $D^{(n)}$. Some of the important considerations are: (a) whether all groups that are treated get treated in the same time period or not, (b) whether

groups that are treated remain treated or whether they may go back to being untreated, and (c) whether the treatment $D_{g,t}$ is binary, continuous, or multi-valued. For our purpose here, we discuss only two designs that involve a binary treatment. The first one, that we termed the basic design, is one with no variation in treatment timing. The second one, that is known as the staggered design, is one where units exhibit variation in treatment timing but where treatment is an absorbing state. More complex designs only exacerbate some of the issues that we illustrate with the staggered design.

13.3.1 Basic DiD Design

We start with the simplest case, where all groups that are treated get treated in the same time period, and remain treated thereafter.

Definition 13.1 (Basic Design) *The treatment assignment takes the form* $D_{g,t} = I\{t \geq t_g^*\}$*, where* $t_g^* \in \{t^*, \infty\}$ *for all* $g \in G$*, for some* $t^* \geq 2$*, and there exists* $g, g' \in G$ *such that* $t_g^* = t^*$ *and* $t_{g'}^* = \infty$*.*

That is, each group is either treated at the same date t^* or never treated (normalized at ∞ for simplicity). In this design we can then partition $G = G_1 \cup G_0$ and $T = T_1 \cup T_0$ as follows:

$$\begin{aligned} G_1 &= \{g \in G : t_g^* = t^*\} = \text{set of treated groups} \\ G_0 &= \{g \in G : t_g^* = \infty\} = \text{set of control groups} \\ T_1 &= \{t \in T : t \geq t^*\} = \text{set of treated time periods} \\ T_0 &= \{t \in T : t < t^*\} = \text{set of control time periods} \end{aligned}$$

Using the Frisch-Waugh-Lovell (FWL) decomposition one can easily show that the TWFE estimator $\hat{\beta}^{\text{fe}}$ of β^{fe} in (13.8) equals $\hat{\theta}^{\text{did}}$, where

$$\hat{\theta}^{\text{did}} = \frac{1}{|G_1|} \sum_{g \in G_1} \hat{\Delta}_g - \frac{1}{|G_0|} \sum_{g \in G_0} \hat{\Delta}_g \ , \tag{13.9}$$

and

$$\hat{\Delta}_g = \frac{1}{|T_1|} \sum_{t \in T_1} Y_{g,t} - \frac{1}{|T_0|} \sum_{t \in T_0} Y_{g,t} \ . \tag{13.10}$$

The estimator $\hat{\theta}^{\text{did}}$ is the difference between the average change of outcome in the treated groups before and after the treatment and the average change of outcome in the control groups before and after the treatment date. It also

follows immediately that

$$\begin{aligned}
E[\hat{\beta}^{\text{fe}}] &= \frac{1}{|G_1|}\sum_{g\in G_1} E[\hat{\Delta}_g] - \frac{1}{|G_0|}\sum_{g\in G_0} E[\hat{\Delta}_g] \\
&= \frac{1}{|G_1|}\sum_{g\in G_1}\left(\frac{1}{|T_1|}\sum_{t\in T_1} E[Y_{g,t}] - \frac{1}{|T_0|}\sum_{t\in T_0} E[Y_{g,t}]\right) \\
&\quad - \frac{1}{|G_0|}\sum_{g\in G_0}\left(\frac{1}{|T_1|}\sum_{t\in T_1} E[Y_{g,t}] - \frac{1}{|T_0|}\sum_{t\in T_0} E[Y_{g,t}]\right) \ ,
\end{aligned}$$

where we used the fact that the design is conditioned upon. If we now use the separability representation in (13.7), the fact that we can write

$$E[Y_{g,t}] = E[Y_{g,t}(0)] + D_{g,t}E[Y_{g,t}(1) - Y_{g,t}(0)] \ , \tag{13.11}$$

and the following definition of an average treatment effect (ATT),

$$ATT = \frac{1}{|G_1||T_1|}\sum_{g\in G_1}\sum_{t\in T_1} E[Y_{g,t}(1) - Y_{g,t}(0)] \ , \tag{13.12}$$

we obtain that

$$\begin{aligned}
E[\hat{\beta}^{\text{fe}}] &= \frac{1}{|G_1||T_1|}\sum_{g\in G_1}\sum_{t\in T_1} E[Y_{g,t}(1) - Y_{g,t}(0)] \\
&\quad + \frac{1}{|G_1||T_1|}\sum_{g\in G_1}\sum_{t\in T_1} E[Y_{g,t}(0)] - \frac{1}{|G_1||T_0|}\sum_{g\in G_1}\sum_{t\in T_0} E[Y_{g,t}(0)] \\
&\quad - \left(\frac{1}{|G_0||T_1|}\sum_{g\in G_0}\sum_{t\in T_1} E[Y_{g,t}(0)] - \frac{1}{|G_0||T_0|}\sum_{g\in G_0}\sum_{t\in T_0} E[Y_{g,t}(0)]\right) \\
&= ATT + \frac{1}{|G_1||T_1|}\sum_{g\in G_1}\sum_{t\in T_1}(\eta_g + \gamma_t) - \frac{1}{|G_1||T_0|}\sum_{g\in G_1}\sum_{t\in T_0}(\eta_g + \gamma_t) \\
&\quad - \left(\frac{1}{|G_0||T_1|}\sum_{g\in G_0}\sum_{t\in T_1}(\eta_g + \gamma_t) - \frac{1}{|G_0||T_0|}\sum_{g\in G_0}\sum_{t\in T_0}(\eta_g + \gamma_t)\right) \\
&= ATT \ .
\end{aligned}$$

The last equality follows because the group effects and time effects cancel:

$$\begin{aligned}
\frac{1}{|G_1||T_1|}\sum_{g\in G_1}\sum_{t\in T_1}\eta_g - \frac{1}{|G_1||T_0|}\sum_{g\in G_1}\sum_{t\in T_0}\eta_g = 0 \ , \\
\frac{1}{|G_1||T_1|}\sum_{g\in G_1}\sum_{t\in T_1}\gamma_t - \frac{1}{|G_0||T_1|}\sum_{g\in G_0}\sum_{t\in T_1}\gamma_t = 0 \ ,
\end{aligned}$$

with the analogous cancellations for $G_0 \times T_0$ and $G_1 \times T_0$. Thus, the TWFE estimator is unbiased for the ATT in the basic design, where ATT is the average (over treated groups and treated periods) of the average treatment effect on the treated. In particular, notice that it may be the case that $E[Y_{g,t}(1) - Y_{g,t}(0)] \neq E[Y_{g',t}(1) - Y_{g',t}(0)]$ for $g \neq g'$.

13.3.2 Staggered DiD Design

In the basic design of Definition 13.1 the TWFE estimator $\hat{\beta}^{\text{fe}}$ is a simple DiD estimator, and so it identifies the ATT under a parallel-trends assumption. However, such a desirable property does not necessarily translate to more sophisticated designs and, in general, $\hat{\beta}^{\text{fe}}$ may fail to estimate an ATT-like parameter if the treatment effect varies across groups and time periods. While this rather negative result about the TWFE estimator applies broadly whenever we depart from the basic design, for the purpose of this class we restrict attention to the case where the treatment continues to be binary, treatment is an absorbing state, but the timing of treatment varies across groups. This is known as the staggered design.

Definition 13.2 (Staggered Design) *The treatment assignment takes the form* $D_{g,t} = I\{t \geq t_g^*\}$, *with* $\min_{g \in G: t_g^* \geq 2} t_g^* < \max_{g \in G} t_g^*$.

Again, t_g^* is the first date at which group g becomes treated and, once treated, group g remains treated thereafter. If g never becomes treated over the study period, we let $t_g^* > |T|$, e.g., $t_g^* = \infty$. We allow $t_g^* = 1$, which corresponds to a group that is treated in every period of the panel, a so-called *always-treated* group. However, $\min_{g \in G: t_g^* \geq 2} t_g^* < \max_{g \in G} t_g^*$ requires that among groups that are untreated at period 1, not all groups get treated at the same period. If that condition fails, then we wouldn't have a valid control group for the groups that become treated after period 1 and the two-way fixed effects regression would not be well defined. The only difference between the staggered design and the basic design is that in the staggered design, the date when treated groups get treated can vary across groups: there can be variation in treatment timing.

Let

$$TE_{g,t} := E[Y_{g,t}(1) - Y_{g,t}(0)]$$

denote the expected treatment effect in cell (g,t) of moving the treatment from 0 to 1. The decomposition below is most easily understood as an FWL decomposition of $\hat{\beta}^{\text{fe}}$: regress $D_{g,t}$ on group and time fixed effects, get residuals $\hat{V}_{g,t}$, and then reweight outcomes by these residuals. Theorem 1 in De Chaisemartin and d'Haultfoeuille [2020] shows that under the assumptions we have

made so far, $\hat{\beta}^{\text{fe}}$ is unbiased for a weighted sum of the $TE_{g,t}$. That is,

$$E[\hat{\beta}^{\text{fe}}] = \frac{1}{|G||T|} \sum_{g \in G} \sum_{t \in T} W_{g,t} TE_{g,t} \ , \tag{13.13}$$

where

$$W_{g,t} := \frac{\hat{V}_{g,t} D_{g,t}}{\frac{1}{|G||T|} \sum_{g' \in G} \sum_{t' \in T} \hat{V}_{g',t'} D_{g',t'}} \tag{13.14}$$

and $\hat{V}_{g,t}$ denotes the residual from a regression of $D_{g,t}$ on group and period fixed effects. The proof of this result is left as a problem exercise, see Problem 13.2.

Note that it follows from the definition that the weights $W_{g,t}$ are normalized so that their average over cells is equal to 1, i.e.,

$$\frac{1}{|G||T|} \sum_{g \in G} \sum_{t \in T} W_{g,t} = 1 \ .$$

However, it is not true that the weights are non-negative independently of the design of the treatment assignment $D^{(n)}$. The result in the previous section implies that in the basic design $W_{g,t} = 1$ for all $(g,t) \in G_1 \times T_1$ and so, not only are the weights non-negative, but they are also all equal to one which leads to $\hat{\beta}^{\text{fe}}$ being unbiased for the ATT. In general, however, the result in (13.13) implies that $\hat{\beta}^{\text{fe}}$ identifies a weighted average of $TE_{g,t}$ with weights that may potentially be *negative*.

In order to appreciate this point further in the context of the staggered design, note that one can show that

$$\hat{V}_{g,t} = D_{g,t} - \bar{D}_{g,\cdot} - \bar{D}_{\cdot,t} + \bar{D} \tag{13.15}$$

where $\bar{D}_{g,\cdot}$ is the average treatment of group g across periods, $\bar{D}_{\cdot,t}$ is the average treatment at period t across groups, and $\bar{D}$ is the average treatment across groups and periods. The expression in (13.15) offers some insights as to when the weights $W_{g,t}$ would be expected to be positive or negative. In particular, some of the weights $W_{g,t}$ may be negative, if there are (g,t) cells such that

$$1 + \bar{D} < \bar{D}_{g,\cdot} + \bar{D}_{\cdot,t} \ .$$

Groups whose average treatment $\bar{D}_{g,\cdot}$ is high are likely to have negative weights. In the staggered design, $\bar{D}_{g,\cdot} = (|T| - t_g^* + 1)/|T|$, so groups whose average treatment is the highest are those for which t_g^* is the lowest, namely the groups that become treated early. Always treated groups are such that $\bar{D}_{g,\cdot} = 1$, so for them $\hat{V}_{g,t} = \bar{D} - \bar{D}_{\cdot,t}$. As $\bar{D}_{\cdot,t}$ is weakly increasing in t when treatment is an absorbing state, $\bar{D}_{\cdot,|T|} > \bar{D}$, so if there are always treated groups, their treatment effect at the last period is always weighted negatively by $\hat{\beta}^{\text{fe}}$. In terms of time periods, the later periods of the panel are more likely to lead to negative weights due to $\bar{D}_{\cdot,t}$ being weakly increasing in t.

The weights are all positive if and only if $1 + \bar{D} > \bar{D}_{g,\cdot} + \bar{D}_{\cdot,t}$ for all (g,t). Accordingly, all the weights are likely to be positive when there is no group that is treated most of the time, and no time period where most groups are treated.

The inability of the TWFE estimator to identify the ATT also applies to other standard regression-based estimators. For example, the first difference estimator that arises from a regression of $Y_{g,t} - Y_{g,t-1}$, the outcome's first difference, on $D_{g,t} - D_{g,t-1}$, the treatment's first difference, and period fixed effects, can also be decomposed as a weighted sum of $TE_{g,t}$ with weights $W^{\text{fd}}_{g,t}$ that differ from those in (13.14) but that also sum to one and that may also be negative. If we let this estimator be denoted by $\hat{\beta}^{\text{fd}}$, the result can be written as:

$$E[\hat{\beta}^{\text{fd}}] = \frac{1}{|G||T|} \sum_{g \in G} \sum_{t \in T} W^{\text{fd}}_{g,t} TE_{g,t} \ . \tag{13.16}$$

13.4 Empirical Illustration

The point of the following illustration is to show that negative weights are not just a theoretical possibility. In real datasets, the regression estimand can place substantial negative weight on some treatment effects, which complicates the interpretation of an otherwise familiar regression coefficient.

Gentzkow et al. [2011] use an 1868-1928 US county-level panel data set to test the conjecture that newspapers encourage citizens to participate more in democratic institutions. For that purpose, they let $Y_{g,t}$ denote the turnout rate in county g and the presidential election that took place in year t, they let $D_{g,t}$ denote the number of newspapers in county g and year t, and they run a regression of the change in turnout in county g between two elections on the change of the county's number of newspapers (including state-year fixed effects). That is, they compute $\hat{\beta}^{\text{fd}}$.

To accommodate potentially non-binary treatments, we redefine the treatment effect $TE_{g,t}$. Under Assumptions 13.1-13.2, for all (g,t) such that $D_{g,t} \neq 0$, let

$$TE_{g,t} = \frac{E[Y_{g,t}(D_{g,t}) - Y_{g,t}(0)]}{D_{g,t}} \ .$$

That is, $TE_{g,t}$ denotes the expected effect in cell (g,t) of moving the treatment from 0 to $D_{g,t}$, scaled by $D_{g,t}$. In other words, $TE_{g,t}$ is the slope of (g,t)'s potential outcome function, from 0 to its *actual* treatment. In Gentzkow et al. (2011), this is the difference between the actual turnout rate in county g and year t and its counterfactual turnout rate without any newspaper, divided by its number of newspapers. Thus, $TE_{g,t}$ can be interpreted as an effect per newspaper.

The results from the paper are that $\hat{\beta}^{\text{fd}} = 0.0026$: one more newspaper

increases turnout by 0.26 percentage points. The coefficient is significant at all conventional levels (with a clustered standard error (at county level) of 0.0009). Using the `twowayfeweights` Stata package, it is easy to find that under parallel trends, $\hat{\beta}^{\text{fd}}$ estimates a weighted sum of the effects of newspapers on turnout in 9,752 county×election-year cells, where 5,286 effects are weighted positively while 4,466 are weighted negatively, and where negative weights sum to -1.37. Accordingly, $\hat{\beta}^{\text{fd}}$ is far from estimating a convex combination of treatment effects. The weights are negatively correlated with the election year, and the correlation is significant: $\hat{\beta}^{\text{fd}}$ is more likely to upweight newspapers' effects in early elections, and to downweight or weight negatively newspapers' effects in late elections.

13.5 Final Remarks

I would like to thank Clément de Chaisemartin for sharing useful material that made it possible to write these notes. A large fraction of these notes are based on his book in progress [de Chaisemartin and d'Haultfoeuille, 2023], the original paper [De Chaisemartin and d'Haultfoeuille, 2020], and related work on modern DiD methods, including event-study and robustness approaches that we discuss in the next DiD chapter [Sun and Abraham, 2021, Borusyak et al., 2023, Rambachan and Roth, 2023, Roth et al., 2022]. Many other important references are available in those papers.

13.6 Problems

Problem 13.1 *Prove* (13.7).

Problem 13.2 *Prove* (13.13).

Problem 13.3 *Prove* (13.15).

Problem 13.4 *Consider the simple* 2×2 *setup from the beginning of the chapter: two groups* $g \in \{s, n\}$ *and two periods* $t \in \{1, 2\}$*, with* $D_{s,2} = 1$ *and* $D_{g,t} = 0$ *otherwise. Show that the OLS coefficient on* $D_{g,t}$ *in the TWFE regression*

$$Y_{g,t} = \tilde{\eta}_g + \tilde{\gamma}_t + \beta^{\text{fe}} D_{g,t} + U_{g,t}$$

fitted to the four observations $\{(Y_{g,t}, D_{g,t}) : g \in \{s, n\}, t \in \{1, 2\}\}$ *equals exactly*

$$\hat{\beta}^{\text{fe}} = (Y_{s,2} - Y_{s,1}) - (Y_{n,2} - Y_{n,1}) \ ,$$

the simple DiD estimator. Hint: apply Frisch-Waugh-Lovell to partial out the group and time fixed effects. The residual of $D_{g,t}$ after this partialling out is $\frac{1}{4}$ at $(s,2)$, $-\frac{1}{4}$ at $(s,1)$ and $(n,2)$, and $\frac{1}{4}$ at $(n,1)$.

Problem 13.5 *Consider a staggered design with three groups $G = \{a, b, c\}$ and four periods $T = \{1, 2, 3, 4\}$. Group a is treated starting at $t_a^* = 2$, group b at $t_b^* = 4$, and group c is never treated.*

(a) Compute the residuals $\hat{V}_{g,t}$ from (13.15) for every (g,t) and verify that $\hat{V}_{a,4} < 0$ while $\hat{V}_{g,t} \geq 0$ at every other treated cell.

(b) Use part (a) to find the weights $W_{g,t}$ in (13.14) for the four treated cells $(a,2)$, $(a,3)$, $(a,4)$, and $(b,4)$.

(c) Construct values of $TE_{g,t}$ that are strictly positive *in every treated cell and yet make $E[\hat{\beta}^{\text{fe}}] < 0$. (*Hint: load enough weight onto $TE_{a,4}$.*) Explain in one paragraph why this is a problem for the empirical interpretation of TWFE estimates in staggered designs.*

Bibliography

K. Borusyak, X. Jaravel, and J. Spiess. Revisiting event study designs: Robust and efficient estimation, 2023.

D. Card and A. B. Krueger. Minimum wages and employment: a case study of the fast-food industry in new jersey and pennsylvania. *The American Economic Review*, 84:772–793, 1994.

C. De Chaisemartin and X. d'Haultfoeuille. Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, 110(9): 2964–96, 2020.

C. de Chaisemartin and X. d'Haultfoeuille. Difference-in-differences for simple and complex natural experiments. November 2023. Available at SSRN: https://ssrn.com/abstract=4487202.

M. Gentzkow, J. M. Shapiro, and M. Sinkinson. The effect of newspaper entry and exit on electoral politics. *American Economic Review*, 101(7): 2980–3018, 2011.

A. Rambachan and J. Roth. A More Credible Approach to Parallel Trends. *The Review of Economic Studies*, 90(5):2555–2591, 02 2023. doi: 10.1093/restud/rdad018. URL https://doi.org/10.1093/restud/rdad018.

J. Roth, P. H. Sant'Anna, A. Bilinski, and J. Poe. What's trending in difference-in-differences? a synthesis of the recent econometrics literature. *arXiv preprint arXiv:2201.01194*, 2022.

L. Sun and S. Abraham. Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, 225(2):175–199, 2021.

14

More Difference in Differences

Having covered the basic elements of difference in differences, including the properties of the two way fixed effects (TWFE) estimator in the basic and staggered designs, today we shift attention to event studies and their implications for testing hypotheses about parallel trends. We also discuss ways to relax the parallel trends assumptions to obtain partial identification results, and then introduce synthetic controls as an alternative approach for settings with one or few treated groups.

14.1 Event Studies in the Basic Design

An event study is a difference-in-differences (DiD) design that intends to be "dynamic" and is typically used to display a range of treatment effects that are present before and after the policy change. As a result, event studies are often used as a tool to test for the no-anticipation assumption and parallel trends. While there exist a variety of ways in which specifications of this type are typically implemented, the most standard specification takes the following form:

$$Y_{g,t} = \eta_g + \gamma_t + \sum_{\ell \in L} \beta_\ell^{\text{fe}} I\{t = t_g^* + \ell\} + U_{g,t} \ , \tag{14.1}$$

where $L = \{-\underline{\ell}, \ldots, \bar{\ell}\} \setminus \{-1\}$, $\underline{\ell} \geq 0$ and $\bar{\ell} \geq 0$ are the numbers of included "leads" and "lags" of the event indicator, respectively. The first lead, $\ell = -1$ is often excluded as a normalization, while the coefficients on the other leads (if present) are interpreted as measures of "pre-trends". For $\ell \geq 0$, β_ℓ^{fe} is supposed to capture the cumulative effect of $\ell + 1$ treatment periods. For $\ell < -1$, β_ℓ^{fe} is supposed to be a placebo coefficient testing the parallel trends assumption, by comparing the outcome trends of groups that will and will not start receiving the treatment in $|\ell|$ periods. The regression in (14.1) is often referred to as a TWFE event study regression, and researchers often plot the estimated coefficients $\{\hat{\beta}_\ell^{\text{fe}} : \ell \in L\}$ on a so-called ES graph, with $\ell = t - t_g^*$, the relative time to treatment onset for the treated groups, on the x-axis. Sometimes the regression in (14.1) includes the additional term $\beta_+ I\{t \geq t_g^* + \bar{\ell}\}$ to capture longer horizons binned together but, for simplicity, we ignore such terms here. It is also important to mention that if there are no never-treated

units, the coefficients $\{\beta_\ell^{\text{fe}} : \ell \in L\}$ are not point-identified. In particular, for any $\kappa \in \mathbf{R}$, $\{\beta_\ell^{\text{fe}} + \kappa(\ell+1) : \ell \in L\}$ is observationally equivalent after the fixed effect coefficients are appropriately modified. Without a never-treated baseline, event-study coefficients and fixed effects are not separately pinned down, so software normalizations can affect the reported path of coefficients. The problem may be important in practice, as statistical packages may resolve this collinearity by dropping an arbitrary unit or period indicator. See Borusyak et al. [2023] for details.

The interpretation of the coefficients associated with event study regressions, as well as the properties of the formal tests for parallel trends, depend crucially on the type of DiD design. While there are many DiD designs in practice, for the purpose of this class we only pay attention to two: the basic design and the staggered design. We start with the basic design, where the treatment is binary and there is no variation in treatment timing:

Definition 14.1 (Basic Design) *The treatment assignment takes the form $D_{g,t} = I\{t \geq t_g^*\}$, where $t_g^* \in \{t^*, \infty\}$ for all $g \in G$, $t^* \geq 2$, and there exists $g, g' \in G$ such that $t_g^* = t^*$ and $t_{g'}^* = \infty$.*

Let $\{\hat{\beta}_\ell^{\text{fe}} : \ell \in L\}$ be the LS estimators of $\{\beta_\ell^{\text{fe}} : \ell \neq -1\}$ in (14.1). In the basic design it is not difficult to show that

$$\hat{\beta}_\ell^{\text{fe}} = \frac{1}{|G_1|} \sum_{g \in G_1} (Y_{g,t^*+\ell} - Y_{g,t^*-1}) - \frac{1}{|G_0|} \sum_{g \in G_0} (Y_{g,t^*+\ell} - Y_{g,t^*-1}) \ , \quad (14.2)$$

a simple DiD comparing the $t^* - 1$ to $t^* + \ell$ outcome evolution in treatment and control groups. This follows from Frisch–Waugh–Lovell: after partialling out group and time fixed effects, the coefficient on the event-time dummy compares the treated-control change from $t^* - 1$ to $t^* + \ell$. All DiDs are relative to $t^* - 1$, the period prior to the treatment onset for the treated groups, and the omitted time period in (14.1). For $\ell < -1$, $\hat{\beta}_\ell^{\text{fe}}$ is often referred to as a pre-trend or placebo estimator. Intuitively, it assesses if before treatment onset, treatment and control groups were on parallel trends.

Let 0_k and 1_k denote k-dimensional vectors of zeros and ones, respectively. For any $\ell \in \{0, \ldots, |T| - t^*\}$ define

$$ATT_\ell := \frac{1}{|G_1|} \sum_{g \in G_1} E[Y_{g,t^*+\ell}(0_{t^*-1}, 1_{\ell+1}) - Y_{g,t^*+\ell}(0_{t^*+\ell})] \ , \quad (14.3)$$

the average treatment effect of having been treated for $\ell + 1$ periods, across all treated groups and at period $t^* + \ell$.

In the basic design, and under Assumptions 13.1 (no anticipation) and 13.4 (general parallel trends), the coefficients $\hat{\beta}_\ell^{\text{fe}}$ have two important properties. First, for all $\ell \in \{0, \ldots, |T| - t^*\}$ it follows that

$$E[\hat{\beta}_\ell^{\text{fe}}] = ATT_\ell \ . \quad (14.4)$$

Second, whenever $t^* \geq 3$, it follows that for all $\ell \in \{-t^*+1, \ldots, -2\}$,

$$E[\hat{\beta}^{\text{fe}}_\ell] = 0 \ . \tag{14.5}$$

Equation (14.5) is an implication of Assumptions 13.1 and 13.4, and so if $t^* \geq 3$ these assumptions can be tested via the following null and alternative,

$$\begin{aligned} &H_0 : E[\hat{\beta}^{\text{fe}}_\ell] = 0 \text{ for all } \ell \in \{-t^*+1, \ldots, -2\} \\ &\quad \text{vs} \\ &H_1 : E[\hat{\beta}^{\text{fe}}_\ell] \neq 0 \text{ for some } \ell \in \{-t^*+1, \ldots, -2\} \ . \end{aligned}$$

However, testing the hypothesis above has two salient issues; both of which are discussed by Roth [2022]. First, it turns out that tests of parallel trends are often under-powered, which means that the probability of rejecting when H_1 is true is small even when the magnitude of the pre-trends is significant enough to introduce important biases in the estimators of ATT_ℓ. Roth [2022] shows, using DGPs calibrated to published applications, that conventional pre-trend tests can have low power against violations of parallel trends that would generate economically meaningful bias. Second, if used as a pre-test whereby the researcher first tests for the presence of parallel trends and then decides whether to continue with the DiD analysis or not, then tests of parallel trends lead to pre-testing problems (as it is typically the case with pre-testing in general).

To be more concrete about the second issue, suppose that researchers test H_0 above as a way to decide whether the analysis should be continued, or to decide which specification should be reported: say, adding control variables, adding group-specific linear trends, change the definition of the control group, etc. Assume such specification search process continues until the parallel-trends test does not reject H_0. The result of this exercise is that the vector of estimated effects being reported, $\{\hat{\beta}^{\text{fe}}_\ell : \ell \geq 0\}$, is actually *conditional* on values of the pre-trends coefficients $\{\hat{\beta}^{\text{fe}}_\ell : \ell \leq -2\}$ such that the pre-trends test is not rejected. The problem with this approach is that, if we let E_{no} denote the event that the pre-test does not reject H_0 above, for $\ell \geq 0$

$$E[\hat{\beta}^{\text{fe}}_\ell \mid E_{\text{no}}] \neq E[\hat{\beta}^{\text{fe}}_\ell] \ . \tag{14.6}$$

That is, pre-testing could lead to a bias that is distinct from the potential bias that may come from differential trends. The implications go beyond this and also affect inference, where the approximate distribution of $\hat{\beta}^{\text{fe}}_\ell$ conditional on E_{no} deviates from the traditional normal approximations. The combination of low power and pre-testing problems raises serious concerns to the practice of testing for parallel trends as a tool for specification searches.

14.1.1 Application to Benzarti and Carloni (2019)

In July of 2009, France reduced its VAT on sit-down restaurants from 19.6 to 5.5 percent. Benzarti and Carloni (2019) analyze the impact of this change

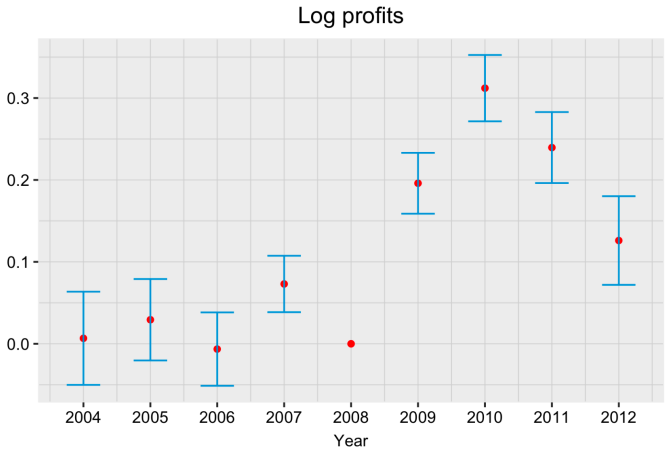


FIGURE 14.1: Event-study estimates of the effect of the VAT cut on firms' profits: Benzarti and Carloni (2019).

using the regression in (14.1), with the control group defined as other market services firms that were not affected by the VAT change. Figure 14.1 shows one of the event-study graphs in their paper, with the outcome defined as the log of firms' profits.

The figure shows evidence of a violation of Assumptions 1 and 3, with a significant positive pre-trend coefficient the year before the reform. However, the estimated effects are much larger than that pre-trend coefficient so it does not seem that pre-trends can account for the entirety of the estimated effects. According to the point estimates, the reform increased restaurants' profits by around 20% the year of the reform, by 30% the following year, by 25% two years after, and by 15% three years after.

14.2 Event Studies in Staggered Designs

The TWFE event study regression in (14.1) in the basic design leads to estimators $\{\hat{\beta}_\ell^{\text{fe}} : \ell \in L\}$ satisfying (14.4)-(14.5). These two properties facilitate the interpretation of these estimators, both for $\ell \leq -2$ and $\ell \geq 0$, as in Figure 14.1. In many settings, however, groups enter treatment at different points in time and so we now consider the properties of $\{\hat{\beta}_\ell^{\text{fe}} : \ell \in L\}$ in the staggered design. To recap, this design is defined as follows.

Definition 14.2 (Staggered Design) *The treatment assignment takes the form* $D_{g,t} = I\{t \geq t_g^*\}$, *with* $\min_{g \in G: t_g^* \geq 2} t_g^* < \max_{g \in G} t_g^*$.

That is, the treatment continues to be binary but the timing of treatment varies across groups. Importantly, once treated, a group remains treated so there is no "off" switch. When it comes to event studies in the context of staggered designs, researchers often run the regression we defined in (14.1) in order to estimate dynamic effects and test the no-anticipation and parallel-trends assumptions. There are multiple variants of this regression, but the specification we consider here is quite standard and the main implications we discuss below apply to all of them.

In the previous class we learned that the properties of the "static" TWFE estimator were drastically different between the basic and staggered designs. As a result, it should probably not be a surprise that TWFE event-study regressions are also not particularly robust to heterogeneous effects and often lead to weighted averages of specific type of treatment effects with weights that may potentially be *negative*. In order to present these results formally, we start by describing the relevant "treatment effects" in this context.

Let 0_k and 1_k denote k-dimensional vectors of zeros and ones, respectively. For all g such that $t_g^* \leq |T|$, and for $\ell \in \{0, \ldots, |T| - t_g^*\}$, let

$$TE_{g,\ell} = E[Y_{g,t_g^*+\ell}(0_{t_g^*-1}, 1_{\ell+1}) - Y_{g,t_g^*+\ell}(0_{t_g^*+\ell})] \ . \tag{14.7}$$

In words, $TE_{g,\ell}$ is the expected treatment effect, in group g at period $t_g^* + \ell$ of having been treated rather than untreated from period t_g^* to $t_g^* + \ell$, namely for $\ell + 1$ periods. It follows that $\hat{\beta}_\ell^{\text{fe}}$ in (14.1) satisfies, for $\ell \in \{0, \ldots, \bar{\ell}\}$,

$$E[\hat{\beta}_\ell^{\text{fe}}] = \sum_{g \in G_{(\ell)}} W_{g,\ell} TE_{g,\ell} + \sum_{\ell' \neq \ell} \sum_{g \in G_{(\ell')}} W_{g,\ell'}^\star TE_{g,\ell'} \ , \tag{14.8}$$

where

$$G_{(\ell)} := \{g \in G : t_g^* \leq |T| - \ell\} \ ,$$

and $W_{g,\ell}$ and $W_{g,\ell}^\star$ are weights such that $\sum_{g \in G_{(\ell)}} W_{g,\ell} = 1$ and $\sum_{g \in G_{(\ell)}} W_{g,\ell}^\star = 0$.

The first term in the right-hand side of (14.8) is a weighted sum across groups of the cumulative effect of $\ell + 1$ treatment periods. As in previous derivations, these weights add up to 1 but may be negative. In this sense this first summation resembles that in the decomposition of $\hat{\beta}^{\text{fe}}$ that we have seen before. The second term in the right-hand side of (14.8) is a weighted sum (across $\ell' \neq \ell$ and groups) of the cumulative effect of $\ell' + 1$ treatment periods in group g, with weights summing up to 0. This second summation was not present in the decomposition of $\hat{\beta}^{\text{fe}}$ and it implies that $\hat{\beta}_\ell^{\text{fe}}$, which is supposed to estimate the cumulative effect of $\ell + 1$ treatment periods, may include contamination from the effects of $\ell' + 1$ treatment periods.

Decompositions for the TWFE event study estimators $\{\hat{\beta}_\ell^{\text{fe}} : \ell \in L\}$ for

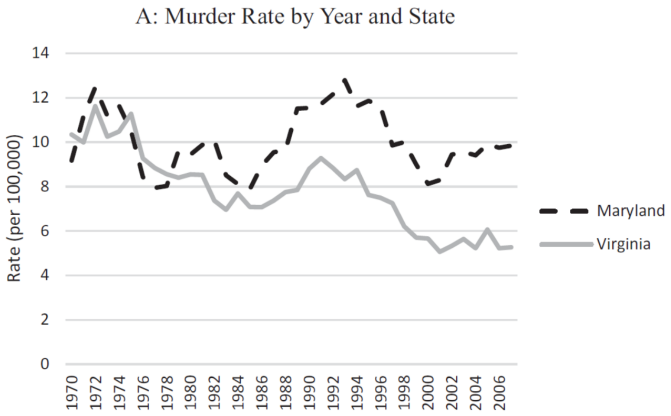


FIGURE 14.2: Murder rate by year and state: Manski and Pepper (2018)

$\ell < -1$, like the one in (14.8), can be used to show that when treatment effects are expected to be heterogeneous, TWFE event-study regressions cannot be used to test the no-anticipation and parallel-trends assumptions (see Sun and Abraham [2021]). This follows from the second term in the decomposition for $E[\hat{\beta}_\ell^{\text{fe}}]$. When $\ell < -1$ the first term measures differential trends between groups that will and will not start receiving the treatment in $|\ell|$ periods. But the second term is a weighted sum, across $\ell' \geq 0$ and groups, and so the expectation of $\hat{\beta}_\ell^{\text{fe}}$ may differ from zero even if parallel trends holds, and it may be equal to zero even if parallel trends fails. It follows immediately that TWFE event-study regressions in staggered designs should be used with care.

14.3 Relaxing Parallel Trends

Due to the skepticism on the validity of parallel trends in many applications, as well as the problems associated with testing for such assumptions, Manski & Pepper (2018) Manski and Pepper [2018] propose to relax the parallel trends assumption by what they call bounded-variation assumptions. The essence of the approach is to give up on point identification results, opt for a partial identification analysis of the treatment effects, and consider a spectrum of assumptions with varying identifying power. Concretely, Manski & Pepper aim to study average treatment effects of right-to-carry (RTC) laws on crime rates in the US. For example, in 1989 Virginia adopted a RTC law whereas Maryland (among other states), did not. Figure 14.2 displays murder rates by year for these two states.

Manski & Pepper (2018) consider an array of alternative assumptions that would allow for (partial) identification of the ATT, but for the purposes of this class we focus on the one that pertains to Assumption 13.3; i.e., parallel trends. To present their idea, let us assume that Assumptions 13.1-13.2 hold, and that $|T| = t^* = 3$. In this section we also go back to the basic design, where the treatment is binary and there is no variation in treatment timing in order to isolate the argument from additional complications that arise in staggered adoption designs. Concretely, the authors propose to replace parallel trends by the following weaker assumption:

Assumption 14.1 (Bounded Variation) *There is a positive real number* Δ *such that*

$$\left| E\left[\frac{1}{|G_1|}\sum_{g\in G_1}(Y_{g,3}(0) - Y_{g,2}(0))\right] - E\left[\frac{1}{|G_0|}\sum_{g\in G_0}(Y_{g,3}(0) - Y_{g,2}(0))\right]\right| \leq \Delta \ .$$

This assumption allows treated and control groups to experience differential trends, but requires the differential trend to be bounded in absolute value by some constant Δ. Rambachan and Roth [2023] build upon Manski & Pepper (2018) and consider an alternative to the bounded variation assumption. In particular, they consider the following stationary differential trends assumption.

Assumption 14.2 (Stationary differential trends) *There is a positive real number* Δ *such that*

$$\left| E\left[\frac{1}{|G_1|}\sum_{g\in G_1}(Y_{g,3}(0) - Y_{g,2}(0))\right] - E\left[\frac{1}{|G_0|}\sum_{g\in G_0}(Y_{g,3}(0) - Y_{g,2}(0))\right]\right| \leq$$

$$\Delta \times \left| E\left[\frac{1}{|G_1|}\sum_{g\in G_1}(Y_{g,2}(0) - Y_{g,1}(0))\right] - E\left[\frac{1}{|G_0|}\sum_{g\in G_0}(Y_{g,2}(0) - Y_{g,1}(0))\right]\right| \ .$$

This assumption requires that the period-2-to-3 differential trend be bounded in absolute value by some constant Δ times the period-1-to-2 differential trend. In other words, differential trends cannot vary "too much" from period to period, where "too much" is controlled by Δ. Under bounded variation, Δ is measured in outcome units; under stationary differential trends, Δ is a relative bound that scales the pre-treatment differential trend.

Under any of these assumptions, and restricting attention to the basic design with binary treatment and no treatment variation, the ATT is *partially* identified and can be characterized by either a linear programming problem or a set of moment inequalities. It follows that inference on the ATT can be done by exploiting tools from these literatures, a topic that is beyond the scope of this class but that is often covered in advanced inference courses.

The partial identification approach leads to bounds on the parameter of interest given a value of Δ; a feature that needs to be determined by the researcher. Oftentimes, researchers would consider a range of values for Δ and present results for each of these. Alternatively, researchers may opt to follow what is sometimes known as a "break-down" approach. For concreteness, suppose that the estimated value of $\hat{\beta}_{\ell}^{\text{fe}}$ is positive and significant. Under parallel trends, researchers would conclude that the treatment has a positive effect. The break-down approach would ask the following question: what would be the minimal relaxation to parallel trends we would need in order for 0 to be included in the confidence interval for the ATT? That is, what is the lowest value of Δ in Assumption 14.2, denoted by Δ^*, such that 0 belongs to the confidence interval of ATT. If $\Delta^* = 5$, that means that even under differential trends five times larger from period 2 to 3 than from period 1 to 2, one can still conclude that the treatment had a positive effect: the researcher's conclusion is robust to allowing for differential trends. On the other hand, $\Delta^* = 0.2$ means that differential trends five times smaller from period 2 to 3 than from period 1 to 2 are enough for the researcher's conclusion to break down, thus suggesting that results are not robust to plausible differential trends. There are R and Stata packages to compute such a value of Δ. A useful starting point is to ask what values of Δ are consistent with the observed pre-trends, and to bracket the analysis around that range.

In Figure 14.1, there is a significant pre-trend coefficient, but the estimated effects are much larger than the pre-trends. Consistent with that, Figure 14.3 below adapts the right panel of Figure 5 in Rambachan and Roth [2023] to show that results concerning ATT_1, the ATE in 2009, the year of the reform, are robust to violations of parallel trends: Δ^* is just below 2 for that parameter. The same panel also shows that results concerning ATT, the average treatment effect across all years after 2009, are less robust: for that parameter, Δ^* is just below 1.

14.4 Synthetic Controls

Empirical applications with one or few treated groups and many control groups are ubiquitous in economics. The DiD approach as described above in essence treats all control groups as being of equal quality as a control group. This may not be true and so the researcher may want to somehow weight the controls in order to give more importance to those controls that seem "better" for the given treated group. This is basically the idea of the synthetic control method, originally proposed by Abadie et al. [2010] (ADH). The application in their paper is the effect of California's tobacco control program on state-wide smoking rates. During the time period in question, there were 38 states in the US that did not implement such programs. Rather than just using a standard

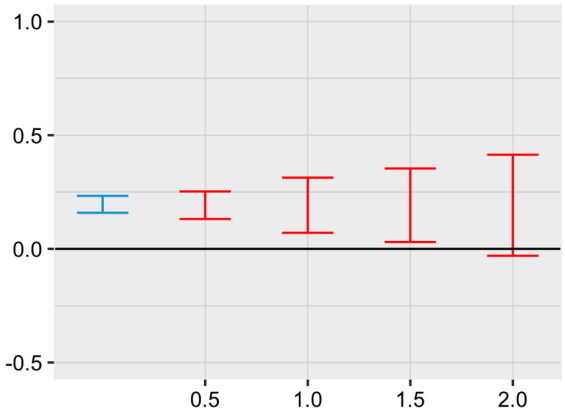


FIGURE 14.3: Confidence intervals for the effect of the VAT cut on firms' profits in 2009, under Assumption 14.2 and varying Δ.

DiD analysis - which effectively treats each state as being of equal quality as a control group - ADH propose choosing a weighted average of the potential controls. Of course, choosing a suitable control group or groups is often done informally, including matching on pre-treatment predictors. ADH formalize the procedure by optimally choosing weights, and they propose methods for inference. In addition, it turns out that synthetic control methods relax parallel trends by allowing for a multiplicative factor structure that we describe below.

Consider the simple case in Section 13.1.2 with only two time periods, one treated group $G_1 = \{s\}$, but potentially many control groups, $|G_0| > 1$. Synthetic control methods allow the model for potential outcomes to be more flexible than the standard additive structure implied by parallel trend assumptions in DiD models. The key departure from the additive DiD model is that the group component and the time component can interact, so untreated trends need not be parallel across groups. To be concrete, in what follows assume that

$$Y_{g,t}(0) = \eta_g \gamma_t + U_{g,t} \ , \tag{14.9}$$

where $E[U_{g,t}] = 0$, so that now the time effect and the group effect interact with each other. Note that common trends do not hold in this model since,

$$E[Y_{g,t}(0) - Y_{g,t-1}(0)] = \eta_g(\gamma_t - \gamma_{t-1}) \ ,$$

which clearly depends on g. Comparing $Y_{s,2}$ and $Y_{g,2}$ for any $g \in G_0$ delivers

$$E[Y_{s,2} - Y_{g,2}] = E[Y_{s,2}(1) - Y_{g,2}(0)] = \theta_{\text{att}} + \gamma_2(\eta_s - \eta_g) \ ,$$

and so this approach does not identify θ in the presence of persistent group

differences. The idea behind synthetic controls is to construct the so-called *synthetic control*

$$\tilde{Y}_{s,2}(0) = \sum_{g \in G_0} w_g Y_{g,2} \ ,$$

by appropriately choosing the weights $\{w_g : g \in G_0, w_g \geq 0, \sum_{g \in G_0} w_g = 1\}$. In order for this idea to work, it must be the case that $E[Y_{s,2}(0)] = E[\tilde{Y}_{s,2}(0)]$ so that $E\left[Y_{s,2} - \tilde{Y}_{s,2}(0)\right] = \theta_{\text{att}}$. Now, for a given set of weights, this approach delivers

$$E\left[Y_{s,2} - \tilde{Y}_{s,2}(0)\right] = E\left[Y_{s,2} - \sum_{g \in G_0} w_g Y_{g,2}\right] = \theta_{\text{att}} + \gamma_2 \left(\eta_s - \sum_{g \in G_0} w_g \eta_g\right) \ .$$

It follows that this approach identifies θ if we could choose the weights in a way such that

$$\eta_s = \sum_{g \in G_0} w_g \eta_g \ . \tag{14.10}$$

This is, however, not feasible as we do not observe the group effects η_g. The main result in Abadie et al. [2010] can be stated for the example in this section as follows: suppose that there exists weights $\{w_g^* : g \in G_0, w_g^* \geq 0, \sum_{g \in G_0} w_g^* = 1\}$ such that

$$Y_{s,1} = \sum_{g \in G_0} w_g^* Y_{g,1} \ . \tag{14.11}$$

If we construct the synthetic control using these optimal weights w_g^*,

$$\tilde{Y}_{s,2}(0) = \sum_{g \in G_0} w_g^* Y_{g,2} \ ,$$

then it follows that $E\left[Y_{s,2} - \tilde{Y}_{s,2}(0)\right] = \theta_{\text{att}}$.

Proving this result in the context of our example is straightforward if we make the further simplifying assumption that the weights $\{w_g^* : g \in G_0, w_g^* \geq 0, \sum_{g \in G_0} w_g^* = 1\}$ are non-stochastic. First, note that by (14.11) we get that

$$\eta_s \gamma_1 + U_{s,1} = \sum_{g \in G_0} w_g^* \eta_g \gamma_1 + \sum_{g \in G_0} w_g^* U_{g,1} \ ,$$

so that

$$\gamma_1 \left(\eta_s - \sum_{g \in G_0} w_g^* \eta_g\right) = - \sum_{g \in G_0} w_g^* (U_{s,1} - U_{g,1}) \ . \tag{14.12}$$

Next note that

$$\begin{aligned} Y_{s,2} - \tilde{Y}_{s,2}(0) &= \theta_{\text{att}} + \eta_s\gamma_2 + U_{s,2} - \sum_{g\in G_0} w_g^*(\eta_g\gamma_2 + U_{g,2}) \\ &= \theta_{\text{att}} + \gamma_2\left(\eta_s - \sum_{g\in G_0} w_g^*\eta_g\right) + \sum_{g\in G_0} w_g^*(U_{s,2} - U_{g,2}) \\ &= \theta_{\text{att}} - \frac{\gamma_2}{\gamma_1}\sum_{g\in G_0} w_g^*(U_{s,1} - U_{g,1}) + \sum_{g\in G_0} w_g^*(U_{s,2} - U_{g,2}) \ , \end{aligned}$$

where we used (14.12) in the third equality. The result follows from $E[U_{g,t}] = 0$ for all (g,t).

We then get the weights by "matching" the observed outcomes of the treated group and the control groups in the period before the policy change. In practice, $Y_{s,1}$ may not lie in the convex hull of $\{Y_{g,1} : g \in G_0\}$ and so the method relies on minimizing the distance between $Y_{s,1}$ and $\sum_{g\in G_0} w_g Y_{g,1}$. Abadie et al. [2010] provide some formal arguments around these issues, and in particular require that the number of pre-treatment periods $|T_0| \to \infty$ and that $U_{g,t}$ is independent across g and t.

The basic idea can be extended in the presence of covariates X_g that are not (or would not be) affected by the policy change. In this case, the weights would be chosen to minimize the distance between

$$(Y_{s,1}, X_s) \text{ and } \sum_{g\in G_0} w_g(Y_{g,1}, X_g) \ .$$

The optimal weights - which differ depending on how we define distance - produce the synthetic control whose pre-intervention outcome and predictors of post-intervention outcome are "closest". Abadie et al. [2010] propose permutation methods for inference. The usual idea is to compare the treated unit's post-treatment gap to placebo gaps obtained by pretending, one at a time, that each control unit was treated.

14.5 Concluding Remarks

I would like to thank Clément de Chaisemartin for sharing useful material that made it possible to write these notes. A large fraction of these notes are based on his book in progress, de Chaisemartin and d'Haultfoeuille [2023], and several related papers, like Abadie et al. [2010], De Chaisemartin and d'Haultfoeuille [2020], Manski and Pepper [2018], Rambachan and Roth [2023], Sun and Abraham [2021], Borusyak et al. [2023]. Many other important references are available in those papers.

14.6 Problems

Problem 14.1 *Prove* (14.2).

Problem 14.2 *Can $\ell \mapsto ATT_\ell$ in* (14.3) *be used to determine if past treatments affect the outcome? For instance, if $ATT_2 > ATT_1 > 0$, can we conclude that the first treatment lag has an effect on the current outcome?*

Problem 14.3 *Prove* (14.4) *and* (14.5).

Problem 14.4 *Consider a staggered design with three groups $G = \{a, b, c\}$ and four periods $T = \{1, 2, 3, 4\}$. Group a is treated starting at $t_a^* = 2$, group b at $t_b^* = 3$, and group c is never treated. Assume the no-anticipation, no-dynamics, and parallel-trends assumptions hold, and that all treatment effects $TE_{g,\ell}$ are well-defined.*

(a) Compute the OLS coefficient $\hat{\beta}_0^{\text{fe}}$ in the TWFE event-study regression (14.1) *with leads $\underline{\ell} = 2$ and lags $\bar{\ell} = 2$, and $\ell = -1$ omitted. Express $\hat{\beta}_0^{\text{fe}}$ as a linear combination of the observed $Y_{g,t}$.*

(b) Using your expression from (a), write $E[\hat{\beta}_0^{\text{fe}}]$ as a weighted sum of the population treatment effects $TE_{g,\ell}$, $\ell \in \{0, 1, 2\}$, $g \in \{a, b\}$. Identify the coefficient on $TE_{a,1}$ and explain why it is not zero: in the language of (14.8)*, this is the contamination from $\ell' \neq \ell$ effects.*

(c) Construct a specification of treatment effects in which $TE_{a,0} = TE_{b,0} > 0$, $TE_{a,1} > 0$, $TE_{a,2} > 0$, and yet $E[\hat{\beta}_0^{\text{fe}}] < TE_{a,0} = TE_{b,0}$. Explain in one paragraph what this implies for interpreting event-study point estimates in staggered designs.

Hint: in (a), apply Frisch–Waugh–Lovell to partial out the group and period fixed effects from each event-time indicator $I\{t = t_g^* + \ell\}$; the residuals are small in absolute value and easy to write down explicitly.

Problem 14.5 *Consider the basic design with $|T| = t^* = 3$, and assume Assumptions 13.1–13.2. For simplicity, use the absolute bounded-variation version in Assumption 14.1, so that Δ is measured in outcome units. Suppose the researcher observes pre- and post-treatment outcomes that imply the DiD estimate $\hat{\theta}^{\text{did}} = \hat{\beta}_0^{\text{fe}} = \hat{\tau} > 0$ with estimated standard error $\hat{\sigma}_\tau$.*

(a) Under Assumption 14.1 with parameter $\Delta \geq 0$, characterize the identified set for θ_{att} as a function of Δ and $\hat{\tau}$. Give the half-width of the identified set as Δ grows.

(b) Define the break-down value Δ^* as the smallest $\Delta \geq 0$ such that 0 lies in the 95% confidence interval for θ_{att}. Express Δ^* as a function of $\hat{\tau}$ and $\hat{\sigma}_\tau$, ignoring the Imbens–Manski correction and using the standard $\hat{\tau} \pm 1.96\,\hat{\sigma}_\tau$ form.

(c) Now suppose the researcher also observes a pre-trend estimate $\hat{\beta}^{\text{fe}}_{-2}$. True or false, and justify: if $|\hat{\beta}^{\text{fe}}_{-2}| > \Delta^*$, the conclusion that treatment had a positive effect is no longer robust to bounded-variation deviations from parallel trends, holding the magnitude of the bound at the observed pre-trend.

Bibliography

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.

K. Borusyak, X. Jaravel, and J. Spiess. Revisiting event study designs: Robust and efficient estimation, 2023.

C. De Chaisemartin and X. d'Haultfoeuille. Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, 110(9): 2964–96, 2020.

C. de Chaisemartin and X. d'Haultfoeuille. Difference-in-differences for simple and complex natural experiments. November 2023. Available at SSRN: https://ssrn.com/abstract=4487202.

C. F. Manski and J. V. Pepper. How do right-to-carry laws affect crime rates? coping with ambiguity using bounded-variation assumptions. *Review of Economics and Statistics*, 100(2):232–244, 2018.

A. Rambachan and J. Roth. A More Credible Approach to Parallel Trends. *The Review of Economic Studies*, 90(5):2555–2591, 02 2023. doi: 10.1093/restud/rdad018. URL https://doi.org/10.1093/restud/rdad018.

J. Roth. Pre-test with caution: Event-study estimates after testing for parallel trends. *American Economic Review: Insights (forthcoming)*, 2022.

L. Sun and S. Abraham. Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, 225(2):175–199, 2021.

15

RDD

We have discussed analyzing the causal effect of a treatment on outcomes of interest under various experimental settings. As many interventions of interest to economists cannot be randomly assigned, upcoming lectures will study research designs for rigorous study of non-experimental interventions. Today we will study the Regression Discontinuity (RD) design, where all units have a score, and a treatment is assigned to those units whose value of the score exceeds a known cutoff or threshold, and not assigned to those units whose value of the score is below the cutoff. The key feature of the design is that the probability of receiving the treatment changes abruptly at the known threshold. If units are unable to perfectly "sort" around this threshold, the discontinuous change in this probability can be used to learn about the local causal effect of the treatment on an outcome of interest, because units with scores barely below the cutoff can be used as a comparison group for units with scores barely above it.

We will focus on the canonical sharp RD design that has the following features: (i) the score is continuously distributed and has only one dimension, (ii) there is only one cutoff, and (iii) compliance with treatment assignment is perfect, i.e., all units with score equal to or greater than the cutoff actually receive the treatment, and all units with score below the cutoff fail to receive the treatment and instead receive the control condition. We will discuss the required assumptions and interpretation for identifying the RD treatment effects, the appropriate methods for estimation and inference, the graphical illustration of the design, and the main validation exercises used in applications.

15.1 Sharp RDD: Identification

Consider the following setting: there are n units, indexed by $i = 1, 2, \ldots, n$, each unit has a score X_i. c is a known cutoff: units with $X_i \geq c$ are assigned to the treatment condition, and units with $X_i < c$ are assigned to the control condition. Thus the treatment assignment can be defined as $A_i = I\{X_i \geq c\}$. In other words, the conditional probability of receiving treatment given the

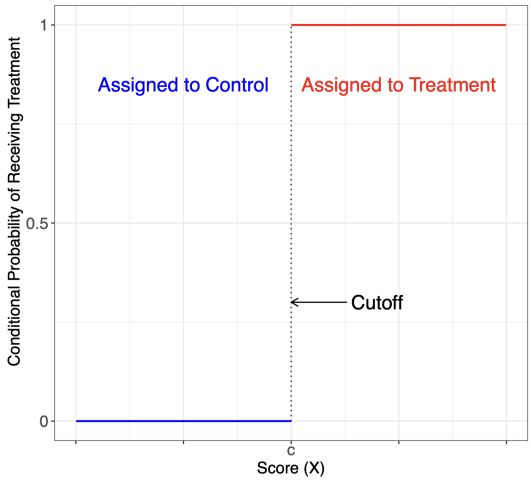


FIGURE 15.1: Conditional Probability of Receiving Treatment in the Sharp RD Design

score, $P\{A_i = 1 \mid X_i = x\}$, changes from 0 to 1 at $X_i = c$, as illustrated by Figure 15.1.

We adopt the potential outcomes framework and assume that each unit has two potential outcomes, $Y_i(1)$ and $Y_i(0)$, corresponding to the outcomes that would be observed under the treatment or control conditions. The observed outcome is

$$Y_i = (1 - A_i) \cdot Y_i(0) + A_i \cdot Y_i(1) = \begin{cases} Y_i(0) & \text{if } X_i < c \\ Y_i(1) & \text{if } X_i \geq c \end{cases}.$$

The fundamental problem of causal inference arises because we only observe the outcome under control, $Y_i(0)$, for units whose score is below the cutoff, and we only observe the outcome under treatment, $Y_i(1)$, for units above the cutoff. As shown in Figure 15.2, the regression function $E[Y_i(1) \mid X_i]$ is observed for values of the score to the right of the cutoff, represented by the solid red line. However, to the left of the cutoff, all units are untreated, so $E[Y_i(1) \mid X_i]$ remains unobserved (represented by the dashed red line). A similar issue arises for $E[Y_i(0) \mid X_i]$, which is observed for values of the score to the left of the cutoff (solid blue line), but unobserved for $X_i \geq c$ (dashed blue line). As a result, the observed average outcome, given the score, is

$$E[Y_i \mid X_i] = \begin{cases} E[Y_i(0) \mid X_i] & \text{if } X_i < c \\ E[Y_i(1) \mid X_i] & \text{if } X_i \geq c \end{cases}. \tag{15.1}$$

The average treatment effect at a given value of the score,

$$E[Y_i(1) \mid X_i = x] - E[Y_i(0) \mid X_i = x] \ ,$$

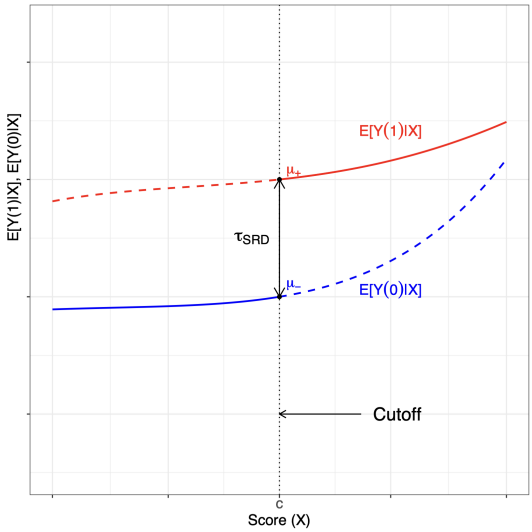


FIGURE 15.2: RD Treatment Effect in Sharp RD Design

is the vertical distance between the two regression curves at that value. This distance cannot be directly estimated because we never observe both curves for the same value of x. However, a special situation occurs at the cutoff c: this is the only point at which we "almost" observe both curves. To see this, we imagine having units with score exactly equal to c, and units with score barely below c (that is, with score $c-\epsilon$ for a small and positive ϵ). The former units would receive treatment, and the latter would receive control. Yet if the values of the average potential outcomes at c are not abruptly different from their values at points near c, the units with $X_i = c$ and $X_i = c - \epsilon$ would be very similar except for their treatment status, and we could approximately calculate the vertical distance at c using observed outcomes. This motivates the Sharp RD treatment effect

$$\theta_{\text{srd}} \equiv E\left[Y_i(1) - Y_i(0) \mid X_i = c\right] = E\left[Y_i(1) \mid X_i = c\right] - E\left[Y_i(0) \mid X_i = c\right].$$

The assumption of comparability between units with very similar values of the score but on opposite sides of the cutoff can be formalized as the continuity of the regression functions $E\left[Y_i(1) \mid X_i = x\right]$ and $E\left[Y_i(0) \mid X_i = x\right]$ at $x = c$. Under this assumption, we have

$$\begin{aligned} E\left[Y_i(1) - Y_i(0) \mid X_i = c\right] &= E\left[Y_i(1) \mid X_i = c\right] - E\left[Y_i(0) \mid X_i = c\right] \\ &= \lim_{x \downarrow c} E\left[Y_i(1) \mid X_i = x\right] - \lim_{x \uparrow c} E\left[Y_i(0) \mid X_i = x\right] \\ &= \lim_{x \downarrow c} E\left[Y_i \mid X_i = x\right] - \lim_{x \uparrow c} E\left[Y_i \mid X_i = x\right], \quad (15.2) \end{aligned}$$

where the second equality is due to the continuity assumption, and the last

equality is due to (15.1). The last expression is the difference between the limits of the treated and control average observed outcomes as the score converges to the cutoff, which is identifiable from the data.

The Sharp RD parameter presented above can be interpreted as causal in the sense that it captures the average difference in potential outcomes under treatment versus control. However, this average difference is calculated at a single point on the support of the score, and as a result is local in nature. In the absence of assumptions about the global shape of the regression functions, the effect recovered by the RD design is only the average effect of treatment at the cutoff, interpreted as a limiting comparison of units whose score values are just below and just above c.

15.2 Estimation via Local Linear Regression

By (15.2), estimating the RD treatment effect requires approximating the regression function $E[Y_i \mid X_i = x]$ at $x = c$ from both sides of the cutoff. We have seen in 480-2 that such a regression function can be estimated by various non-parametric methods. In particular, local polynomial estimators are predominantly used in RD analysis due to their robustness at boundary points. The RD-specific point is that we estimate two one-sided limits at the cutoff and take their difference. Local polynomial estimation consists of the following basic steps.

1. Choose a polynomial order p, a kernel function $K(\cdot)$, and a bandwidth h.
2. For observations above the cutoff, fit a weighted least squares regression of the outcome Y_i on a constant and $(X_i - c), (X_i - c)^2, \ldots, (X_i - c)^p$, with weight $K\left(\frac{X_i - c}{h}\right)$ for each observation. The estimated intercept from this regression, $\hat{\mu}_n^+$, is an estimate of $\mu^+ = E[Y_i(1) \mid X_i = c]$.
3. For observations below the cutoff, do the same weighted least squares regression. The estimated intercept from this regression, $\hat{\mu}_n^-$, is an estimate of $\mu^- = E[Y_i(0) \mid X_i = c]$.
4. Calculate the Sharp RD point estimate:

$$\hat{\theta}_{\text{srd}} = \hat{\mu}_n^+ - \hat{\mu}_n^- \tag{15.3}$$

A graphical representation of local polynomial RD point estimation is given in Figure 15.3, where a polynomial of order one ($p = 1$) is fit; observations outside bandwidth h are not used in the estimation.

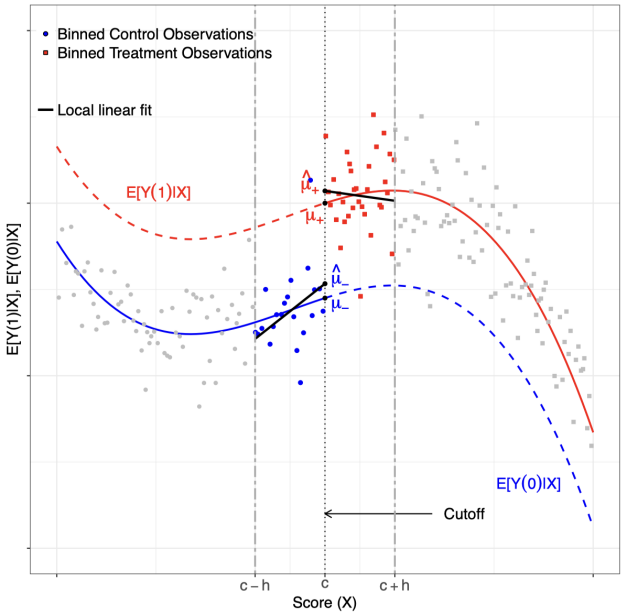


FIGURE 15.3: RD Estimation with Local Polynomial

15.2.1 Bandwidth Choice

The implementation of the local polynomial approach requires the choice of three main ingredients: the kernel function, the order of the polynomial, and the bandwidth. We refer to 480-2 for the first two, and devote our discussion to bandwidth choice.

The most popular approach in practice seeks to minimize the MSE of the local polynomial RD point estimator, $\hat{\theta}_{\text{srd}}$, given a choice of polynomial order and kernel function. Since the MSE of an estimator is the sum of its squared bias and its variance, this approach effectively chooses h to optimize a bias-variance trade-off. The following discussions assume a common bandwidth h on the left and right of the cutoff, but can be extended to allow for different bandwidths. The general form of the approximate (conditional) MSE for the RD treatment effect is

$$MSE\left(\hat{\theta}_{\text{srd}}\right) = h^{2(p+1)}\mathcal{B}^2 + \frac{1}{nh}\mathcal{V}.$$

The unknown constants involved in $\mathcal{B}$ are the $(p+1)$th derivatives of the regression functions $E\left[Y(0) \mid X = x\right]$ and $E\left[Y(1) \mid X = x\right]$:

$$\lim_{x \downarrow c} \frac{d^{p+1} E\left[Y(1) \mid X = x\right]}{dx^{p+1}} \quad \text{and} \quad \lim_{x \uparrow c} \frac{d^{p+1} E\left[Y(0) \mid X = x\right]}{dx^{p+1}},$$

which are related to the "curvature" of the unknown regression functions for treatment and control units, respectively.

The unknown constants involved in $\mathcal{V}$ are the ratios between the conditional variance of the outcome given the score and the density of the score at the cutoff, for treatment and control units:

$$\frac{\lim_{x\downarrow c} \operatorname{Var}\left[Y_i(1) \mid X_i = x\right]}{f_X(c)} \quad \text{and} \quad \frac{\lim_{x\uparrow c} \operatorname{Var}\left[Y_i(0) \mid X_i = x\right]}{f_X(c)}.$$

The MSE-optimal bandwidth choice minimizes the MSE approximation:

$$h_{\mathrm{MSE}} = \arg\min_{h>0} \left(h^{2(p+1)} \mathcal{B}^2 + \frac{1}{nh} \mathcal{V} \right) = \left(\frac{\mathcal{V}}{2(p+1)\mathcal{B}^2} \right)^{1/(2p+3)} n^{-1/(2p+3)}.$$

This formula formally incorporates the bias-variance trade-off mentioned above. It follows that the MSE-optimal bandwidth increases with $\mathcal{V}$ and decreases with $\mathcal{B}$. In other words, a larger asymptotic variance will lead to a larger MSE-optimal bandwidth; this is intuitive, as a larger bandwidth will include more observations in the estimation and thus reduce the variance of the resulting point estimator. In contrast, a larger asymptotic bias will lead to a smaller bandwidth, as a smaller bandwidth will reduce the approximation error and reduce the bias of the resulting point estimator.

In practice, the optimal bandwidth selectors described above are implemented by constructing preliminary plug-in estimates of the unknown quantities. For example, the constant $\mathcal{B}$ is estimated by forming preliminary "curvature" estimates, which are constructed using a local polynomial of order $q \geq p+1$ with bandwidth b, not necessarily equal to h. The resulting estimator for $\mathcal{B}$ is denoted by $\hat{\mathcal{B}}$. Similarly, an estimator $\hat{\mathcal{V}}$ for $\mathcal{V}$ is usually constructed using the asymptotic variance of the estimates on the left and right of the cutoff. Natural choices are some version of heteroskedasticity-consistent standard error formulas or modifications thereof allowing for clustered data. Given these ingredients, the MSE-optimal bandwidth choice is

$$\hat{h}_{\mathrm{MSE}} = \left(\frac{\hat{\mathcal{V}}}{2(p+1)\hat{\mathcal{B}}^2} \right)^{1/(2p+3)} n^{-1/(2p+3)}.$$

15.3 Inference

The MSE-optimal bandwidth discussed previously results in an RD point estimator that is both consistent and optimal in an MSE sense, but poses a problem for inference. The challenge is that the chosen bandwidth is not small enough to remove the leading bias term in the standard distributional approximations used to conduct statistical inference. Formally, $\hat{\theta}_{\mathrm{srd}}$ has an approximate large-sample distribution

$$\frac{\hat{\theta}_{\mathrm{srd}} - \theta_{\mathrm{srd}} - \mathscr{B}}{\sqrt{\mathcal{V}}} \approx N(0,1),$$

where $\mathscr{B}$ and $\mathcal{V}$ are, respectively, the asymptotic bias and variance of the RD local polynomial estimator, discussed previously. Then an asymptotic 95% confidence interval for θ_{srd} is approximately given by

$$\text{CI} = \left[\left(\hat{\theta}_{\text{srd}} - \mathscr{B}\right) \pm 1.96 \cdot \sqrt{\mathcal{V}}\right].$$

We now discuss different strategies that are often employed to make inferences for $\hat{\theta}_{\text{srd}}$ based on the above asymptotic distributional approximation.

15.3.1 Conventional Inference and Undersmoothing

A strategy sometimes found in RD empirical work is to ignore the bias term even when an MSE-optimal bandwidth is used, and construct the following confidence interval:

$$\text{CI}_{\text{c}} = \left[\hat{\theta}_{\text{srd}} \pm 1.96 \cdot \sqrt{\hat{\mathcal{V}}}\right],$$

where $\hat{\mathcal{V}}$ is the estimated asymptotic variance of the RD local polynomial estimator. This leads to invalid inferences in all cases except when the bias term is so small that it can be ignored.

A theoretically sound but ad hoc alternative procedure is to use the same form of confidence interval with a smaller (or "undersmoothed") bandwidth relative to the MSE-optimal one used for construction of the point estimator:

$$\text{CI}_{\text{us}} = \left[\hat{\theta}_{\text{srd}} \pm 1.96 \cdot \sqrt{\hat{\mathcal{V}}}\right].$$

Practically, this procedure involves first selecting the MSE-optimal bandwidth, then selecting a bandwidth smaller than the MSE-optimal choice, and finally constructing the above confidence interval with this smaller bandwidth—note that the latter step requires estimating both a new point estimator and a new standard error with the smaller bandwidth. The theoretical justification is that, for bandwidths sufficiently smaller than the MSE-optimal choice, the bias term will become negligible as $n \to \infty$.

The main drawback of this undersmoothing procedure is that there are no clear and transparent criteria for shrinking the bandwidth below the MSE-optimal value: some researchers might estimate the MSE-optimal choice and divide by two, others may choose to divide by three, etc. Although these procedures can be justified in theory, they are all ad hoc and can result in lack of transparency and specification searching. Moreover, this general strategy leads to a loss of statistical power because a smaller bandwidth results in fewer observations used for estimation and inference.

15.3.2 Standard Bias Correction

Inference could be based on the MSE-optimal bandwidth so long as the induced asymptotic bias is manually estimated and removed from the distributional approximation. This approach, known as bias correction, first estimates

the bias term $\mathscr{B}$ with the estimator $\hat{\mathscr{B}}$ (which in fact is already estimated for implementation of MSE-optimal bandwidth selection), and then constructs confidence intervals that are centered at the bias-corrected point estimate:

$$\mathrm{CI}_{\mathrm{bc}} = \left[\left(\hat{\theta}_{\mathrm{srd}} - \hat{\mathscr{B}}\right) \pm 1.96 \cdot \sqrt{\hat{\mathcal{V}}}\right] .$$

As explained above, the bias term depends on the the $(p+1)$th derivatives of the regression functions at the cutoff. These unknown derivatives can be estimated with a local polynomial of order $q \geq p+1$, which requires another choice of bandwidth, denoted b. Therefore, the RD point estimate $\hat{\theta}_{\mathrm{srd}}$ employs the bandwidth h, while the bias estimate $\hat{\mathscr{B}}$ employs the additional bandwidth b. The ratio $\rho = h/b$ is important, as it relates to the variability of the bias correction estimate relative to the RD point estimator. Standard bias correction methods require $\rho = h/b \to 0$. In particular, note this rules out $\rho = h/b = 1$, that is, standard bias correction does not allow $h = b$.

The bias-corrected confidence intervals allow for a wider range of bandwidths h and, in particular, result in valid inferences when the MSE-optimal bandwidth is used. However, they typically have poor performance in applications because the variability introduced in the bias estimation step is not incorporated in $\hat{\mathcal{V}}$.

15.3.3 Robust Bias Correction

A superior strategy that is both theoretically sound and leads to improved coverage in finite samples is to use robust bias correction for constructing confidence intervals. The robust bias correction approach delivers valid inferences even when the MSE-optimal bandwidth for point estimation is used – no undersmoothing is necessary – and remains valid even when $h = b$.

Like $\mathrm{CI}_{\mathrm{bc}}$, robust bias-corrected confidence intervals adjust the RD point estimator by the estimated bias $\hat{\mathscr{B}}$. The difference is a new estimated asymptotic variance $\hat{\mathcal{V}}_{\mathrm{bc}}$ that, unlike $\hat{\mathcal{V}}$ used in CI_{c}, $\mathrm{CI}_{\mathrm{us}}$, and $\mathrm{CI}_{\mathrm{bc}}$, incorporates the contribution of the bias correction step to the variability of the bias-corrected point estimator. Because $\hat{\mathcal{V}}_{\mathrm{bc}}$ incorporates the extra variability introduced in the bias estimation step, it is larger than $\hat{\mathcal{V}}$. This approach leads to the robust bias-corrected confidence interval:

$$\mathrm{CI}_{\mathrm{rbc}} = \left[\left(\hat{\theta}_{\mathrm{srd}} - \hat{\mathscr{B}}\right) \pm 1.96 \cdot \sqrt{\hat{\mathcal{V}}_{\mathrm{bc}}}\right] .$$

We summarize the differences between the confidence intervals discussed in Table 15.1.

TABLE 15.1: Local Polynomial Confidence Intervals

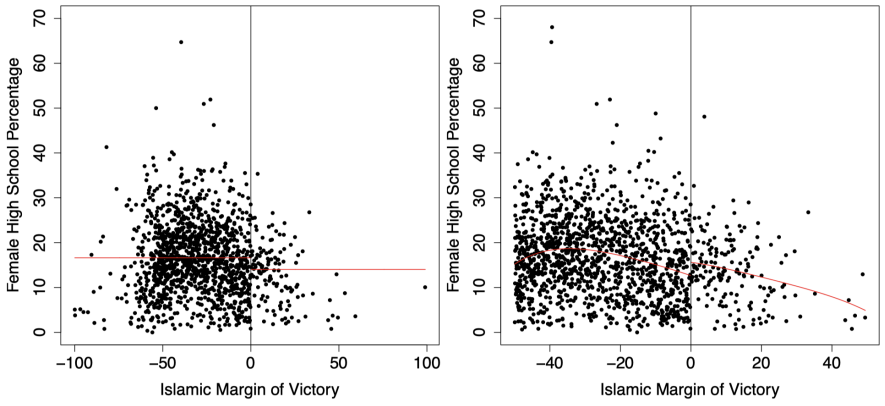
	Centered at	Standard Error
Conventional: CI_c	$\hat{\theta}_\text{srd}$	$\sqrt{\hat{\mathcal{V}}}$
Undersmoothed: CI_us	$\hat{\theta}_\text{srd}$	$\sqrt{\hat{\mathcal{V}}}$
Bias-Corrected: CI_bc	$\hat{\theta}_\text{srd} - \hat{\mathscr{B}}$	$\sqrt{\hat{\mathcal{V}}}$
Robust bias-corrected: CI_rbc	$\hat{\theta}_\text{srd} - \hat{\mathscr{B}}$	$\sqrt{\hat{\mathcal{V}}_\text{bc}}$

15.4 Empirical Example

We now introduce an empirical example that we will employ throughout this and the next lecture, originally analyzed by Meyersson (2014) Meyersson [2014]. Meyersson's original study employs a Sharp RD design, based on close elections in Turkey, to study the impact of having a mayor from an Islamic party on various outcomes. The unit of analysis is the municipality, and the score is the Islamic margin of victory, defined as the difference between the vote percentage obtained by the largest Islamic party, and the vote percentage obtained by the largest secular party opponent. As defined, the Islamic margin of victory can be positive or negative, and the cutoff that determines an Islamic party victory is located at zero. Given this setup, the treatment group consists of municipalities that elected a mayor from an Islamic party in 1994, and the control group consists of municipalities that elected a mayor from a secular party. The outcome we re-analyze is the educational attainment of women who were (potentially) in high school during the period 1994-2000, calculated as the percentage of the cohort of women aged 15 to 20 in 2000 who had completed high school by 2000 according to the 2000 Turkish census. To summarize, the following are variables we will analyse:

- Y: educational attainment of women, measured as the percentage of women aged 15 to 20 in 2000 who had completed high school by 2000.
- X: vote margin obtained by the Islamic party in the 1994 Turkish mayoral elections, measured as the vote percentage obtained by the Islamic party minus the vote percentage obtained by its strongest secular party opponent.
- A: electoral victory of the Islamic party in 1994, equal to 1 if the Islamic party won the mayoral election and 0 otherwise.

The methodological challenge is that municipalities where the support for Islamic parties is high enough to result in the election of an Islamic mayor may differ systematically from municipalities where the support for Islamic parties is lower and results in the election of a secular mayor. If some of the characteristics on which both types of municipalities differ affect (or are correlated



(a) Raw Comparison of Means (b) Local Comparison of Means

FIGURE 15.4: Municipalities with Islamic Mayor versus Municipalities with Secular Mayor

with) the educational outcomes of women, a simple comparison of municipalities with an Islamic versus a secular mayor will be misleading. For example, municipalities where an Islamic mayor wins in 1994 may be more religiously conservative than municipalities where a secular mayor is elected. If religious conservatism affects the educational outcomes of women, the comparison between municipalities controlled by an Islamic versus a secular mayor will not isolate the effect of the Islamic party's control of the local government. Instead, the effect of interest will be contaminated by differences in the degree of religious conservatism between the two groups.

This is illustrated in Figure 15.4, where we plot the percentage of young women who had completed high school by 2000 against the Islamic margin of victory in the 1994 mayoral elections. In Figure 15.4(a), we show the scatter plot of the raw data, superimposing the overall sample mean for each group. The raw comparison reveals a negative average difference: municipalities with an Islamic mayor have, on average, lower educational attainment of women. Figure 15.4(b), shows the scatter plot for the subset of municipalities where the Islamic margin of victory is within 50 percentage points, superimposing a fourth-order polynomial fit separately on either side of the cutoff. Figure 15.4(b) reveals that the negative average effect in Figure 15.4(a) arises because there is an overall negative relationship or slope between Islamic vote percentage and educational attainment of women for the majority of the observations, so that the higher the Islamic margin of victory, the lower the educational attainment of women.

These figures are examples of RD plots, which we discuss in more detail in Appendix C.

15.5 Validation and Extensions

The continuity assumption is not directly testable because it restricts unobserved potential outcome regression functions. However, its observable implications can be investigated. Standard validation and falsification exercises include testing whether the density of the running variable is continuous at the cutoff, estimating RD effects for predetermined covariates and placebo outcomes that should not be affected by the treatment, and checking for discontinuities at placebo cutoffs where the treatment assignment rule does not change. A related robustness check is the donut-hole approach, which drops observations very close to the cutoff to assess sensitivity to the units most likely to have manipulated their score. Appendix C discusses these exercises in the Meyersson application.

Many applications have imperfect compliance: some units with $X_i \geq c$ do not receive treatment, or some units with $X_i < c$ receive treatment anyway. This leads to a Fuzzy RD design, where the local Wald ratio identifies a LATE for compliers at the cutoff under monotonicity and a continuity condition extended to potential treatment status. Appendix C provides the details.

15.6 Concluding Remarks

The material today heavily borrows from two books on RDD: Cattaneo et al. [2019] and Cattaneo et al. [2024], as well as slides shared by Matias Cattaneo. I want to particularly thank Matias for sharing his slides and codes with me. The main inference method discussed today was originally proposed by Calonico et al. [2014].

15.7 Problems

Problem 15.1 *For each statement below, indicate whether it is true or false, and justify your answer in two or three sentences.*

(a) Under continuity of $E[Y_i(1) \mid X_i = x]$ and $E[Y_i(0) \mid X_i = x]$ at $x = c$, the sharp RD parameter θ_{srd} identifies the average treatment effect for the entire population whose score is in some neighborhood of c.

(b) If the density $f_X(c)$ of the running variable at the cutoff is small, the MSE-optimal bandwidth h_{MSE} is small as well.

(c) The conventional confidence interval that ignores the bias is asymptotically valid as long as the bandwidth used to construct $\hat{\theta}_{\text{srd}}$ *is smaller than the MSE-optimal bandwidth.*

(d) In any finite sample, $\text{CI}_{\text{rbc}} \supseteq \text{CI}_{\text{bc}}$.

Bibliography

S. Calonico, M. D. Cattaneo, and R. Titiunik. Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, 82(6): 2295–2326, 2014.

M. D. Cattaneo, N. Idrobo, and R. Titiunik. *A practical introduction to regression discontinuity designs: Foundations*. Cambridge University Press, 2019.

M. D. Cattaneo, N. Idrobo, and R. Titiunik. *A practical introduction to regression discontinuity designs: Extensions*. Cambridge University Press, 2024.

E. Meyersson. Islamic rule and the empowerment of the poor and pious. *Econometrica*, 82(1):229–269, 2014.

Part IV

Inference via Resampling

16

Randomization Tests

In the previous chapters we studied identification and inference in non-experimental designs, including regression discontinuity designs. In this chapter we return to randomized experiments and ask how the randomization itself can be used as the basis for inference. This question has two closely related answers. The first is Fisher's original randomization test, which uses the known treatment assignment mechanism to test a sharp null hypothesis. The second is the more general randomization-test framework of Hoeffding, which uses invariance of the sampling distribution under transformations of the data. Understanding these ideas is useful both for analyzing randomized experiments and for seeing why permutation-style methods can remain informative in more complicated empirical settings.

16.1 The Classical Fisher Framework

Imagine a randomized controlled experiment involving n units, some designated for treatment and others for control. One of the simplest methods to operationalize such an experiment is through 'simple' random sampling (SRS), also known as a Bernoulli randomized experiment (BRE). In a BRE, the treatment assignment vector $\mathbf{A} = \{A_i : 1 \leq i \leq n\}$ is characterized by each A_i being independent and identically distributed Bernoulli random variables with probability π, akin to flipping a coin with probability π. Consequently, if n_1 represents the number of units receiving treatment, then n_1 follows a Binomial(n, π) distribution. Despite its simplicity and intuitive appeal, BREs are not widely adopted in practice due to the inherent risk of generating highly unbalanced assignment outcomes (e.g., $n_1 \approx 0$ or $n_1 \approx n$). This lack of control over n_1 has led to the exploration of alternative treatment assignment methods that mitigate such imbalance issues.

Perhaps the most popular alternative to SRS is complete randomization, also known as a completely randomized experiment (CRE). To define CRE, consider an experiment with n units, with the intended goal of having n_1 receiving the treatment and n_0 receiving the control. Note that n_1 in this case is a number controlled by the researcher (i.e., in an experiment with $n = 100$

units we want to have $n_1 = 30$ in the treatment group). A CRE is defined as follows.

Definition 16.1 *A CRE has the treatment assignment mechanism:*

$$P\{\mathbf{A} = \mathbf{a}\} = 1/\binom{n}{n_1}$$

where $\mathbf{a} = (a_1, \ldots, a_n)$ *satisfies* $\sum_{i=1}^{n} a_i = n_1$ *and* $\sum_{i=1}^{n}(1 - a_i) = n_0$.

An important feature of a CRE, and perhaps the essence of Fisher's framework, is that the potential outcome vector under treatment $\mathbf{Y}(1) = (Y_1(1), \ldots, Y_n(1))$ and the potential outcome vector under control $\mathbf{Y}(0) = (Y_1(0), \ldots, Y_n(0))$ are both viewed as fixed (non-stochastic) numbers. This contrasts with our modeling framework where potential outcomes are random variables, but we will go back to our random framework very soon. For the moment, our most immediate goal is to understand the ideas in the Fisher's framework, which assumes potential outcomes are fixed. Thus, in a CRE, the treatment vector $\mathbf{A}$ is simply one element from a random permutation of n_1 1's and n_0 0's.

In his book Design of Experiments, Fisher (1935) pointed out the following advantages of randomization:

1. It creates comparable treatment and control groups on average.
2. It serves as a "reasoned basis" for statistical inference.

Point 1 is intuitive because the random treatment assignment does not bias toward the treatment or the control. Most people understand point 1 well. Point 2 is more subtle. What Fisher meant is that randomization justifies a statistical test, which is often called the Fisher Randomization Test (FRT). We illustrate the basic idea of the FRT under a CRE below.

16.1.1 The FRT

Fisher (1935) was interested in testing the following null hypothesis:

$$H_{0F} : Y_i(1) = Y_i(0) \quad \text{for all units} \quad i = 1, \ldots, n \ .$$

Rubin (1980) called it the *sharp null hypothesis* in the sense that it can determine all the potential outcomes based on the observed data: $\mathbf{Y}(1) = \mathbf{Y}(0) = \mathbf{Y} = (Y_1, \ldots, Y_n)$.

Conceptually, under H_{0F}, the FRT works for any test statistic $T = T(\mathbf{A}, \mathbf{Y})$, which is a function of the observed data. The observed outcome vector $\mathbf{Y}$ is fixed under H_{0F}, so the only random component in the test statistic T is the treatment vector $\mathbf{A}$. The experimenter determines the distribution of $\mathbf{A}$, which in turn determines the distribution of T under H_{0F}. This is the basis for calculating the p-value as illustrated in Figure 16.1.

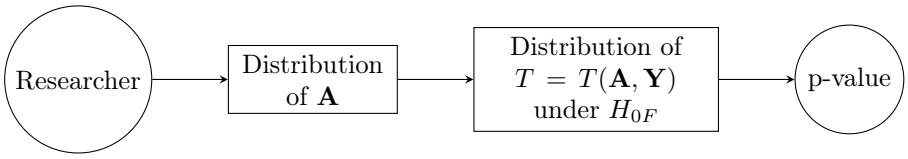


FIGURE 16.1: Illustration of the process from researcher's decision to p-value calculation.

In a CRE, $\mathbf{A}$ is uniform over the set $\{\mathbf{a}_1, \ldots, \mathbf{a}_M\}$ where $M = \binom{n}{n_1}$, and the $\mathbf{a}_m$'s are all possible vectors with n_1 1's and n_0 0's. For instance, with $n = 5$ and $n_1 = 3$, we can enumerate $M = \binom{5}{3} = 10$ vectors as follows:

Listing 16.1: R Code: all M elements

```
> mypermute = function(n, n1){
    M = choose(n, n1)
    treat.ind = combn(n, n1)
    Z = matrix(0, n, M)
  for(m in 1:M){
    treat = treat.ind[, m]
    Z[treat, m] = 1
  }
Z
}
> mypermute(5, 3)
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
[1,]    1    1    1    1    1    1    0    0    0     0
[2,]    1    1    1    0    0    0    1    1    1     0
[3,]    1    0    0    1    1    0    1    1    0     1
[4,]    0    1    0    1    0    1    1    0    1     1
[5,]    0    0    1    0    1    1    0    1    1     1
```

As a consequence, T is uniform over the set (with possible duplications)

$$\{T(\mathbf{a}_1, \mathbf{Y}), \ldots, T(\mathbf{a}_M, \mathbf{Y})\} .$$

That is, the distribution of T is known due to the design of the CRE. We call this distribution of T the *randomization distribution*. Assuming that we choose T so that large values of T provide evidence against the null hypothesis, then we can consider a test $I\{T > cv\}$ where cv is a critical value. In this case, we can use the randomization distribution to compute the desired quantile as a critical value. Alternatively, we can use the following tail probability to measure the extremeness of the test statistic with respect to its randomization distribution:

$$p_{n,\text{frt}} := \frac{1}{M} \sum_{m=1}^{M} I\{T(\mathbf{a}_m, \mathbf{Y}) \geq T(\mathbf{A}, \mathbf{Y})\} , \tag{16.1}$$

which is called the p-value by Fisher. It is the fraction of values T could have taken that are at least as large as the observed value of T. The p-value in (16.1) works for any choice of test statistic and any outcome-generating process. Importantly, it is valid in finite samples in the sense that under H_{0F},

$$P\{p_{n,\text{frt}} \leq u\} \leq u \quad \text{for all } 0 \leq u \leq 1 \ , \tag{16.2}$$

which means that the test that rejects when $p_{n,\text{frt}} \leq \alpha$ is of level α.

In practice, M is often too large (e.g., with $n = 100$, $n_1 = 50$, we have $M > 10^{29}$), and it is computationally infeasible to enumerate all possible values of the treatment vector. This is unproblematic. We often approximate $p_{n,\text{frt}}$ by Monte Carlo. To be more specific, we take simple random draws from the possible values of the treatment vector, or, equivalently, we randomly permute $\mathbf{A}$, and approximate $p_{n,\text{frt}}$ by

$$\hat{p}_{n,\text{frt}} = \frac{1}{B}\left[1 + \sum_{b=1}^{B-1} I\{T(\mathbf{a}_b, \mathbf{Y}) \geq T(\mathbf{A}, \mathbf{Y})\}\right] \ , \tag{16.3}$$

where the $\mathbf{a}_b$'s are the $B-1$ random permutations of $\mathbf{A}$. The p-value $\hat{p}_{n,\text{frt}}$ has Monte Carlo error decreasing fast with an increasing B. Because the calculation of the p-value in (16.3) involves permutations of $\mathbf{A}$, the FRT is also known as the permutation test in the context of the CRE. However, the idea of FRT is more general than the permutation test in more complex experiments.

Listing 16.2: R Code: random permutations

```
treatment = c(1,1,1,0,0)
sample(treatment)
[1] 0 0 1 1 1
sample(treatment)
[1] 0 1 0 1 1
sample(treatment)
[1] 1 0 1 1 0
```

16.1.2 The choice of test statistic

From the above discussion, the FRT generates finite-sample valid p-values for any choice of test statistic. This is a feature of the FRT. However, this feature should not encourage arbitrary choice of the test statistic as the choice affects, among other things, how powerful the resulting test may be. The choice of test statistic is also quite important when we deviate from finite-sample results and consider the asymptotic properties of permutation tests (and, more broadly, randomization tests as defined in the next section).

Perhaps the most canonical choice of test statistic is the absolute value of the difference in means statistic, which is essentially letting T be equal to the

absolute value of $\hat{\theta}_n$ as defined previously in class. To recap, and for $a \in \{0, 1\}$, let

$$\begin{aligned} \bar{Y}_{n,a} &= \frac{1}{n_a} \sum_{1 \leq i \leq n} Y_i I\{A_i = a\} \\ \hat{\sigma}^2_{n,a} &= \frac{1}{n_a} \sum_{1 \leq i \leq n} (Y_i - \bar{Y}_{n,a})^2 I\{A_i = a\} \ . \end{aligned}$$

Setting $T = |\hat{\theta}_n|$ with $\hat{\theta}_n = \bar{Y}_{n,1} - \bar{Y}_{n,0}$ leads to the so-called 'unstudentized' t-test statistic, given that such a test statistic equals the numerator of a 'studentized' t-test statistic. The studentized version would instead be

$$T = \left| \frac{\bar{Y}_{n,1} - \bar{Y}_{n,0}}{\sqrt{\frac{\hat{\sigma}^2_{n,1}}{n_1} + \frac{\hat{\sigma}^2_{n,0}}{n_0}}} \right| \ . \tag{16.4}$$

This test statistic may equivalently be described as the usual t-test (on the coefficient on treatment assignment) in a linear regression of outcomes on treatment assignment with heteroskedasticity-robust standard errors.

It is crucial to recognize that Fisher's framework diverges from our primary framework in two key aspects, both developed in the next section: it treats potential outcomes as fixed, and it tests a sharp pointwise null, $Y_i(1) = Y_i(0)$ for all units i, rather than a null hypothesis on average treatment effects. These features may seem restrictive given our focus on random sampling, randomness beyond the experimental design, and causal parameters characterized by means. However, the basic logic of the Fisher Randomization Test (FRT) remains useful within our adopted framework, which includes a super-population and accounts for sampling-based uncertainty. To see why, the next section distinguishes design-based uncertainty from sampling-based uncertainty.

16.2 Design-based vs Sampling-based Uncertainty

A central distinction in causal inference lies between *design-based* and *model-based* (or sampling-based) inference. In design-based inference, potential outcomes and covariates are treated as fixed properties of the units under study; all randomness comes from the researcher's control over the treatment assignment mechanism. This tradition, rooted in Fisher's framework, views inference as a direct consequence of the deliberate randomization procedure. For example, a permutation test evaluates whether the observed statistic is extreme relative to its distribution generated by considering all possible valid reassignments of treatment to the observed units.

In contrast, model-based inference assumes that data are drawn from a

stochastic process, which we typically call the distribution $P \in \mathbf{P}$. Randomness comes from the sampling process (i.i.d., clustered, Markov chain), and inference relies on probabilistic models that describe this data-generating process - this is the logic behind Hoeffding-style randomization tests that we will discuss in the next section.

One way to summarize the distinction is as follows:

$$
\begin{aligned}
&\text{Design-based:} && \{Y_i(0), Y_i(1) : 1 \leq i \leq N\} \text{ fixed},\\
& && \mathbf{A} \sim \text{known assignment mechanism},\\
& && \theta_N = \frac{1}{N}\sum_{i=1}^{N} \left(Y_i(1) - Y_i(0)\right),\\
&\text{Sampling-based:} && (Y_i(0), Y_i(1), A_i) \overset{i.i.d.}{\sim} P,\\
& && A_i \perp\!\!\!\perp (Y_i(0), Y_i(1)),\\
& && \theta = E_P\left[Y(1) - Y(0)\right].
\end{aligned}
$$

The first framework conditions on the units and asks what would have happened under alternative treatment assignments. The second framework treats the observed units themselves as a random sample and asks what can be learned about the population distribution P.

While both approaches can yield valid inference for different types of causal questions, they differ fundamentally in what is held fixed (the units vs. the data-generating process) and in how uncertainty is conceptualized (arising from treatment assignment vs. arising from sampling variability). Design-based methods strongly emphasize internal validity, as inference is tied to the known randomization procedure, while model-based methods focus on generalizability to the broader population from which the sample was drawn and facilitate extrapolation beyond the observed data, often at the cost of requiring stronger assumptions about the underlying data distribution.

A related distinction is between a *finite population framework* and *super-population framework*. In the finite population approach the data represent a sample of size n extracted without replacement from a finite population of size N. In this case $n \leq N$, and $n = N$ occurs when the entire population is sampled. Causal parameters of interest are therefore defined as averages over the N units; the ATE, for example, is

$$\theta_N = \frac{1}{N}\sum_{i=1}^{N}(Y_i(1) - Y_i(0)) \ .$$

In the super-population approach the data represent an i.i.d. sample of size n from a hypothetical infinite population distribution, referred to as the super-population - note that with an infinite super-population sampling with or without replacement becomes equivalent. Causal parameters of interest are there-

fore defined as expectations over an infinite or hypothetical population distribution P. Although design-based inference is often associated with a focus on finite populations, and model-based inference with super-populations, these pairings are not strict synonyms; it is possible, under certain assumptions, to perform design-based inference to make claims about super-population parameters. Understanding these distinctions is essential for selecting the appropriate inferential tools that align with the research question, the study design, and the target of inference. For recent exposition on these, see [Abadie et al., 2020, 2023] and Imbens and Rubin [2015, Sections 1.12 and 6.7].

16.3 Randomization Tests

In this section, we delve into the finite and large sample behavior of permutation tests and, more broadly, randomization tests. It's important to note that the term 'randomization' takes on a different connotation here compared to its usage in the preceding section. In the previous context, randomization pertained to the pre-data collection process, such as the random assignment of experimental units to treatment or control groups—what we referred to as the treatment assignment mechanism. This randomization facilitated meaningful post-data comparison by enabling the computation of a statistic over permutations of the data. Thus, 'randomization' encompassed both experimental design and data analysis, involving the recomputation of statistics over permutations or randomizations (sometimes termed 'rerandomizations') of the data. It is this latter aspect of randomization that we now extend and generalize. Consequently, the term 'randomization test' denotes tests derived from recalculating a test statistic over transformations (not necessarily permutations) of the data.

Before we describe the general construction of randomization tests, we start our discussion in the context of a simple example that is unrelated to the notion of an experiment.

16.3.1 Motivating example: sign changes

Let $X = (X_1, \ldots, X_{10}) \sim P$ be an i.i.d. sample of size $n = 10$ where each X_i takes values in $\mathbf{R}$, has a finite mean $\theta \in \mathbf{R}$, and has a distribution that is symmetric about θ. For example, $X \sim N(\theta, \sigma^2)$ would satisfy these requirements, but here we do not want to necessarily restrict ourselves to normality assumptions.

Let $\mathbf{P}$ be the collection of all distributions P satisfying these conditions. Consider testing

$$H_0 : \theta = 0 \quad \text{vs} \quad H_1 : \theta \neq 0 \ .$$

We only have 10 observations, so using an asymptotic approximation does not

seem fruitful. At the same time, this is more general than the normal location model where each X_i has distribution $N(\theta, \sigma^2)$, so exploiting normality is not possible.

Let $\bar{X} := n^{-1} \sum_{i=1}^{n} X_i$ and suppose we decide to use the absolute value of $\bar{X}_n$ to test the above hypothesis. Denote this test statistic by $T(X) = |\bar{X}_n|$. The question of interest is as follows: is it possible to construct a critical value $c(1-\alpha)$ that delivers a test $\phi(X) = I\{T(X) > c(1-\alpha)\}$ that is valid in finite samples? By valid here we mean

$$E_P[\phi(X)] \leq \alpha \quad \text{for all } P \in \mathbf{P} \text{ with } \theta = 0 \ .$$

It turns out we can do this by exploiting the fact that the distribution of $X = (X_1, \ldots, X_{10})$ is symmetric about 0 under the null hypothesis; since $X_i \stackrel{d}{=} -X_i$ for all i under the null hypothesis.

To formalize this, let ζ_i be a variable that only takes two values, 1 or -1, for each $i = 1, \ldots, n$. For the moment, these are not random but rather given to us. Now consider a transformation $g = (\zeta_1, \ldots, \zeta_{10})$ of $\mathbf{R}^{10}$ that defines the following mapping

$$(X_1, \ldots, X_{10}) \mapsto gX = (\zeta_1 X_1, \ldots, \zeta_{10} X_{10}) \ . \tag{16.5}$$

Finally, let $\mathbf{G}$ be the $M = 2^{10}$ collection of such transformations. It follows that the random variable X and gX have the *same distribution* under the null hypothesis. What this means is that we can get "new samples" from P by simply applying g to X. We can get a total of $M = 1,024$ samples and use these samples to simulate the distribution of $T(X)$ - following exactly the same intuition we used in the construction of the FRT. This approach leads to a test that is valid in finite samples as the next section shows.

16.3.2 The main result

In this section X denotes the observed sample and P denotes the distribution of the entire sample X (as in the motivating example). Since all results are finite sample in nature, we do not use an index n to denote the sample size and do not index objects by n.

Based on data X taking values in a sample space $\mathcal{X}$, it is desired to test the null hypothesis $H_0 : P \in \mathbf{P}_0$, where P is the true distribution of X and $\mathbf{P}_0$ is a subset of distributions in the space $\mathbf{P}$. Let $\mathbf{G}$ be a finite group of transformations $g : \mathcal{X} \mapsto \mathcal{X}$. The following assumption allows for a general test construction.

Definition 16.2 (Randomization Hypothesis) *Under the null hypothesis, the distribution of X is invariant under the transformations in $\mathbf{G}$; that is, for every $g \in \mathbf{G}$, gX and X have the same distribution whenever $X \sim P \in \mathbf{P}_0$.*

Note that we do not require the alternative hypothesis parameter space to remain invariant under g in $\mathbf{G}$. Only the space $\mathbf{P}_0$ is assumed invariant.

Let $T(X)$ be any real-valued test statistic for testing H_0. Suppose that the group $\mathbf{G}$ has M elements. Given $X = x$, let

$$T^{(1)}(x) \le T^{(2)}(x) \le \cdots \le T^{(M)}(x) \tag{16.6}$$

be ordered values of $T(gX)$ as g varies in $\mathbf{G}$. Fix a nominal level α, $0 < \alpha < 1$, and let k be defined as

$$k = \lceil (1-\alpha)M \rceil \tag{16.7}$$

where $\lceil C \rceil$ denotes the smallest integer greater than or equal to C. Let

$$M^+(x) = \sum_{j=1}^{M} I\{T^{(j)}(x) > T^{(k)}(x)\}$$

$$M^0(x) = \sum_{j=1}^{M} I\{T^{(j)}(x) = T^{(k)}(x)\} \ .$$

Now set

$$a(x) = \frac{M\alpha - M^+(x)}{M^0(x)} \ , \tag{16.8}$$

and define the randomization test as

$$\phi(x) = \begin{cases} 1 & T(x) > T^{(k)}(x) \\ a(x) & T(x) = T^{(k)}(x) \\ 0 & T(x) < T^{(k)}(x) \end{cases} . \tag{16.9}$$

This test is randomized, and since $M^+(x) \le M - k \le M\alpha$ and $M^+(x) + M^0(x) \ge M - k + 1 > M\alpha$, we have $a(x) \in [0, 1)$.

Under the randomization hypothesis, Hoeffding (1952) shows that this construction results in a test of exact level α, and this is true for *any* choice of test statistic $T(X)$. Note that this is possibly a randomized test if $(1-\alpha)M$ is not an integer and there are ties in the ordered values. The randomization at the boundary distributes the rejection probability across observations with $T(X) = T^{(k)}(X)$; without it, a deterministic test generally cannot attain exact size when ties have positive probability. Alternatively, if one prefers not to randomize, the slightly conservative but non-randomized test that rejects when $T(X) > T^{(k)}$, i.e.,

$$\phi^{\mathrm{nr}}(X) = I\{T(X) > T^{(k)}(X)\} \ , \tag{16.10}$$

is level α.

Theorem 16.1 *Suppose that X has distribution P on $\mathcal{X}$ and the problem is to test the null hypothesis $P \in \mathbf{P}_0$. Let $\mathbf{G}$ be a finite group of transformations of $\mathcal{X}$ onto itself. Suppose the randomization hypothesis in Definition 16.2 holds. Given a test statistic $T(X)$, let ϕ be the randomization test as described above. Then, $\phi(X)$ is a similar α level test, i.e.,*

$$E_P[\phi(X)] = \alpha, \ \textit{for all } P \in \mathbf{P}_0 \ .$$

Proof of Theorem 16.1

By construction, for every $x \in \mathcal{X}$,

$$\sum_{g \in \mathbf{G}} \phi(gx) = M^+(x) + a(x)M^0(x) = M\alpha \ ,$$

and so

$$M\alpha = E_P\left[\sum_{g \in \mathbf{G}} \phi(gX)\right] = \sum_{g \in \mathbf{G}} E_P[\phi(gX)] \ .$$

By the randomization hypothesis $E_P[\phi(gX)] = E_P[\phi(X)]$, so that

$$M\alpha = \sum_{g \in \mathbf{G}} E_P[\phi(gX)] = \sum_{g \in \mathbf{G}} E_P[\phi(X)] = ME_P[\phi(X)] \ ,$$

and the result follows.

Remark 16.1 Note that by construction the randomization test not only is of level α for all n, but also "similar", meaning that $E_P[\phi(X)] = \alpha$ for every $P \in \mathbf{P}_0$. ■

In general, one can define a p-value, p_{rt}, of a randomization test by

$$p_{\text{rt}} = \frac{1}{M} \sum_{g \in \mathbf{G}} I\{T(gX) \geq T(X)\} \ . \tag{16.11}$$

It can be shown that p_{rt} satisfies, under the null hypothesis,

$$P\{p_{\text{rt}} \leq u\} \leq u \quad \text{for all } 0 \leq u \leq 1 \ . \tag{16.12}$$

Therefore, the non-randomized test that rejects when $p_{\text{rt}} \leq \alpha$ is level α.

Because $\mathbf{G}$ may be large, one may resort to an approximation to construct the randomization test, for example, by randomly sampling transformations g from $\mathbf{G}$ with or without replacement. In the former case, for example, suppose $g_1, \ldots, g_{B-1}$ are i.i.d. and uniformly distributed on $\mathbf{G}$. Let

$$\hat{p}_{\text{rt}} = \frac{1}{B}\left[1 + \sum_{b=1}^{B-1} I\{T(g_b X) \geq T(X)\}\right] \ . \tag{16.13}$$

Then, it can be shown that, under the null hypothesis,

$$P\{\hat{p}_{\text{rt}} \leq u\} \leq u \quad \text{for all } 0 \leq u \leq 1 \ , \tag{16.14}$$

where this probability reflects variation in both X and the sampling of the g_b.

Example 16.1 (Two sample problem) Suppose that $Y_1, \ldots, Y_m$ are i.i.d.

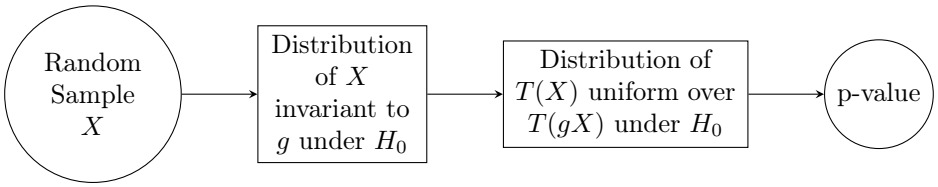


FIGURE 16.2: Intuition behind the validity of a randomization test.

observations from a distribution P_Y and, independently, $Z_1, \ldots, Z_n$ are i.i.d. observations from a distribution P_Z. In other words, we have two samples that are not paired, i.e., Z_1 and Y_1 do not correspond to the same "unit". Here X, the observed data, is given by

$$X = (Y_1, \ldots, Y_m, Z_1, \ldots, Z_n) \ .$$

Consider testing

$$H_0 : P_Y = P_Z \text{ vs } H_1 : P_Y \neq P_Z \ .$$

To describe an appropriate group of transformations $\mathbf{G}$, let $N = m + n$. For $x = (x_1, \ldots, x_N) \in \mathbf{R}^N$, let $gx \in \mathbf{R}^N$ be defined by

$$(x_1, \ldots, x_N) \mapsto gx = (x_{\varsigma(1)}, \ldots, x_{\varsigma(N)}) \ , \tag{16.15}$$

where $(\varsigma(1), \ldots, \varsigma(N))$ is a permutation of $\{1, \ldots, N\}$. Let $\mathbf{G}$ be the collection of all such g, so that $M = N!$. It follows that whenever $P_Y = P_Z$, X and gX have the same distribution.

In essence, each transformation g produces a new data set gx, of which the first m elements are used as the Y sample and the remaining n as the Z sample to recompute the test statistic. Note that, if a test statistic is chosen that is invariant under permutations within each of the Y and Z samples, like

$$T(X) = \bar{Y}_m - \bar{Z}_n \ ,$$

it is enough to consider the $\binom{N}{m}$ transformed data sets obtained by taking m observations from all N as the Y observations and the remaining n as the Z observations (which is equivalent to using a subgroup $\mathbf{G}'$ of $\mathbf{G}$). ■

16.3.3 Permutation tests for treatment effects

Probably the most popular application of randomization tests in economic applications is in the context of randomized controlled experiments, where $\mathbf{G}$ is once again the group of permutations and the problem becomes a type of two sample problem, as the one described in Example 16.1. This application also clarifies the relationship between the FRT and Theorem 16.1. In the FRT, the potential outcome vector is fixed, the null hypothesis is sharp, and the only source of randomness is the design distribution of $\mathbf{A}$. In Theorem

16.1, the sample itself is random and validity follows from invariance of its distribution under transformations in **G**. When the sharp null holds and **G** permutes treatment assignments in a way that preserves (n_1, n_0), the two views produce the same randomization distribution for the statistic. When the null is instead a distributional null or a mean null in a super-population, the theorem helps us see exactly which invariance condition is needed and when it fails.

Suppose that we observe a random sample $\{(Y_1, A_1), \ldots, (Y_n, A_n)\}$ from a randomized controlled trial where

$$Y = Y(1)A + (1 - A)Y(0)$$

is the observed outcome and $A \in \{0, 1\}$ is the exogenous treatment assignment. We assume that $A \perp\!\!\!\perp (Y(1), Y(0))$ but do not make further assumptions about the distribution of A (say, SRS vs CRE). Suppose that we are interested in testing the hypothesis that the distribution Q_0 of $Y(0)$ is the same as the distribution Q_1 of $Y(1)$. That is,

$$H_0 : Q_0 = Q_1 \text{ vs. } H_1 : Q_0 \neq Q_1 \ . \tag{16.16}$$

Note that the observed data are $\{Y_i : A_i = 1\}$ and $\{Y_i : A_i = 0\}$, so the problem is essentially a two-sample problem. Moving an observation from one sample to the other effectively means swapping its treatment assignment A. Under the null hypothesis in (16.16), Y_i and A_i are independent because the marginal distribution of Y_i does not depend on A_i. Therefore,

$$\{(Y_1, A_1), \ldots, (Y_n, A_n)\} \stackrel{d}{=} \{(Y_1, A_{\varsigma(1)}), \ldots, (Y_n, A_{\varsigma(n)})\}$$

for any permutation $(\varsigma(1), \ldots, \varsigma(n))$ of $\{1, \ldots, n\}$. It follows from our general result that a permutation test that permutes individuals from "treatment" to "control" (or from "control" to "treatment") delivers a test that is valid in finite samples.

A few points are worth highlighting on the hypothesis in (16.16). First, the null hypothesis H_0 in (16.16) is weaker than

$$H_0' : Y(1) = Y(0) \quad a.s. \tag{16.17}$$

This is equivalent to the sharp null hypothesis H_{0F} studied in Fisher's framework, except that here it is stated for a super-population framework where potential outcomes are random. Since H_0' implies H_0, one could test this 'sharp' hypothesis using the same construction of a randomization test. Second, the null hypothesis H_0 in (16.16) is stronger than the usual zero 'mean' effect hypothesis,

$$H_0^\star : E[Y(1)] = E[Y(0)] \quad \text{vs} \quad H_1 : E[Y(1)] \neq E[Y(0)] \tag{16.18}$$

The null hypothesis in $H_0^\star$ is often of interest in settings where the researcher

cares about the average treatment effect, since $\theta = E[Y(1)] - E[Y(0)]$. However, it turns out that under $H_0^\star$ there is no group of transformations $\mathbf{G}$ such that $X \stackrel{d}{=} gX$ in general. The reason is that $H_0^\star$ constrains only the means; it still allows, for example, the distribution of Y to differ across treatment states in ways that are not preserved by permuting treatment labels. One may still consider the permutation test that results from considering all possible permutations of the vector of treatment assignment $(A_1, \ldots, A_n)$ and proceed as we previously described. Unfortunately, such an approach does not lead to a valid test and may over-reject in finite samples. The distinction between the null hypothesis in (16.16) and that in (16.18) and their implications on the properties of permutation tests are often ignored in applied research.

16.4 Asymptotic validity of permutation tests

When the goal is to test the null hypothesis $H_0^\star$ in (16.18), which amounts to testing that the ATE is zero, one may still proceed to compute a critical value (or a p-value) via a permutation test instead of relying on the usual asymptotic normality. As we previously mentioned, such a test would not lead to finite sample validity, but one can show that it would be asymptotically valid, as long as the test statistic is carefully chosen. In order to prove this result formally one needs to introduce additional notation and invoke a number of arguments that are beyond the scope of this class. The interested reader should look at [Lehmann and Romano, 2005, Section 15.2.2].

For the purposes of this class, the following two main takeaways in the context of permutation tests for treatment effects are important. First, when the randomization hypothesis is not assumed to hold, say, because the null of interest is $H_0^\star$ and we believe that $Q_0 \neq Q_1$, then the permutation test construction we previously discussed with a difference in means test statistic,

$$T_{\text{unstud}} = \left|\bar{Y}_{n,1} - \bar{Y}_{n,0}\right| \ , \tag{16.19}$$

would lead to an asymptotically valid test (as $n \to \infty$) only if either $n_1 = n_0$ (there are the same number of treated and control units), or if $\text{Var}[Y(0)] = \text{Var}[Y(1)]$ (potential outcomes have equal variances). If the underlying variances differ and $n_1 \neq n_0$, the permutation test based on T_{unstud} will have rejection probability that does not tend to α and may over-reject under the null even in large samples. Second, if instead of using T_{unstud} as the test statistic one resorts to the studentized version

$$T_{\text{stud}} = \frac{\left|\bar{Y}_{n,1} - \bar{Y}_{n,0}\right|}{\sqrt{\frac{\hat{\sigma}^2_{n,1}}{n_1} + \frac{\hat{\sigma}^2_{n,0}}{n_0}}} \ , \tag{16.20}$$

then the permutation test is pointwise consistent in level for testing equality of means, even when the underlying distributions have possibly different

variances and the sample sizes of the treated and control groups differ. Studentization is essential here: T_{stud} is asymptotically pivotal under $H_0^\star$, and the permutation distribution of T_{stud} has the same limiting distribution as the sampling distribution of T_{stud}. The unstudentized statistic is not pivotal when variances differ, so the permutation and sampling distributions need not agree unless the variances or the sample sizes coincide.

Remark 16.2 Randomization tests are often dismissed in applied research due to the belief that the randomization hypothesis is too strong to hold in a real empirical application. This unfortunately does not account for the level of generality in which randomization tests may be asymptotically valid even when the randomization hypothesis does not hold. For example, Bugni et al. [2018] study the asymptotic properties of permutation tests in settings with sophisticated treatment assignment mechanisms, including covariate-adaptive randomization. Moreover, recent developments on the asymptotic properties of randomization tests show that such a construction may be particularly useful in regression models with a fixed and small number of clusters, as shown by Canay et al. [2017]. The approach does not require symmetry in the distribution of X, but rather symmetry in the asymptotic distribution of $\sqrt{n}(\hat{\theta}_n - \theta)$ - which automatically holds when these estimators are asymptotically normal. ■

16.5 Concluding Remarks

The contents of this chapter are based on the notes by Peng Ding Ding [2023] and the book by Lehmann and Romano Lehmann and Romano [2005]. Both of these references are excellent sources for readers who want to know more about randomization and permutation tests.

16.6 Problems

Problem 16.1 *Consider the two-sample setting of Example 16.1. Let* $\{Y_1, \ldots, Y_m\}$ *and* $\{Z_1, \ldots, Z_n\}$ *be independent i.i.d. samples from distributions* P_Y *and* P_Z *on* $\mathbf{R}$*, with means* μ_Y, μ_Z *and variances* $\sigma_Y^2, \sigma_Z^2 \in (0, \infty)$*. Let* $N = m + n$ *and assume* $m/N \to \rho \in (0,1)$ *as* $N \to \infty$*. The objective is to test*

$$H_0^\star : \mu_Y = \mu_Z \quad \textit{vs.} \quad H_1^\star : \mu_Y \neq \mu_Z \ .$$

Let Π *be a permutation of* $\{1, \ldots, N\}$ *drawn uniformly from the symmetric group, independent of the data, and let* $(\bar{Y}_m^\Pi, \bar{Z}_n^\Pi)$ *denote the sample means*

computed after applying Π *to the pooled vector* $(Y_1, \ldots, Y_m, Z_1, \ldots, Z_n)$. *Define*

$$T_{\text{unstud}} = \sqrt{N}\,|\bar{Y}_m - \bar{Z}_n|, \qquad T_{\text{stud}} = \frac{\sqrt{N}\,|\bar{Y}_m - \bar{Z}_n|}{\sqrt{N(\hat{\sigma}^2_{m,Y}/m + \hat{\sigma}^2_{n,Z}/n)}}\,,$$

where $\hat{\sigma}^2_{m,Y}$ *and* $\hat{\sigma}^2_{n,Z}$ *are the corresponding sample variances. Let* T^{Π}_{unstud} *and* T^{Π}_{stud} *denote the same statistics computed on the permuted sample.*

(a) Under $H_0^\star$, *derive the limiting distribution of* $\sqrt{N}(\bar{Y}_m - \bar{Z}_n)$ *under the sampling distribution. Identify the asymptotic variance* V_{samp}.

(b) Conditional on the data, compute the conditional mean and conditional variance of $\sqrt{N}(\bar{Y}^{\Pi}_m - \bar{Z}^{\Pi}_n)$. Hint: *treat the permuted first* m *entries as a simple random sample without replacement from the pooled* N *observations.*

(c) Using a finite-population CLT, state the limiting conditional distribution of $\sqrt{N}(\bar{Y}^{\Pi}_m - \bar{Z}^{\Pi}_n)$ *and the asymptotic conditional variance* V_{perm}.

(d) Show that $V_{\text{samp}} = V_{\text{perm}}$ *if and only if* $\rho = 1/2$ *or* $\sigma_Y^2 = \sigma_Z^2$. *Conclude that the permutation test based on* T_{unstud} *is not asymptotically valid for* $H_0^\star$ *in general.*

(e) Explain why the permutation test based on T_{stud} *is asymptotically valid for* $H_0^\star$ *for any* $\rho \in (0,1)$ *and any* $(\sigma_Y^2, \sigma_Z^2) \in (0,\infty)^2$.

Problem 16.2 *For each of the following statements, decide whether it is true or false. Justify your answer in two to three sentences, and, when the statement is false, identify the minimal additional condition under which it would become true.*

(a) Suppose $T(X)$ *is invariant under the group* $\mathbf{G}$, *that is,* $T(gX) = T(X)$ *for every* $g \in \mathbf{G}$ *and every* X. *Then the randomized test in equation* (16.9) *has rejection probability exactly* α *under every* $P \in \mathbf{P}$, *both inside and outside* $\mathbf{P}_0$.

(b) Under $H_0^\star$: $E[Y(1)] = E[Y(0)]$ *in a CRE with* $A \perp\!\!\!\perp (Y(0), Y(1))$, *the joint distribution of* $(Y_i, A_i)_{i=1}^n$ *is invariant under permutations of* $(A_1, \ldots, A_n)$.

(c) Suppose the randomization hypothesis holds for the strong null H_0 : $Q_0 = Q_1$ *and the group of permutations of* $(A_1, \ldots, A_n)$. *Then the non-randomized permutation test that rejects when* $T_{\text{stud}}(X) > T^{(k)}_{\text{stud}}(X)$ *is of level* α *in finite samples.*

(d) In a CRE with n_1 *treated and* n_0 *control units, the FRT in Section 1 and the two-sample permutation test in Example 16.1 produce the same p-value for any test statistic.*

Bibliography

A. Abadie, S. Athey, G. W. Imbens, and J. M. Wooldridge. Sampling-based versus design-based uncertainty in regression analysis. *Econometrica*, 88(1): 265–296, 2020.

A. Abadie, S. Athey, G. W. Imbens, and J. M. Wooldridge. When should you adjust standard errors for clustering? *The Quarterly Journal of Economics*, 138(1):1–35, 2023.

F. A. Bugni, I. A. Canay, and A. M. Shaikh. Inference under covariate adaptive randomization. *Journal of the American Statistical Association*, 113(524): 1784–1796, 2018. doi: 10.1080/01621459.2017.1375934.

I. A. Canay, J. P. Romano, and A. M. Shaikh. Randomization tests under an approximate symmetry assumption. *Econometrica*, 85(3):1013–1030, May 2017.

P. Ding. A first course in causal inference. arXiv:2305.18793, 2023.

G. W. Imbens and D. B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015.

E. Lehmann and J. P. Romano. *Testing Statistical Hypotheses*. Springer, New York, 3rd edition, 2005.

D. N. Politis, J. P. Romano, and M. Wolf. *Subsampling*. Springer, New York, 1999.

17

The Bootstrap

The bootstrap is a powerful and flexible approach to inference that is quite popular in applied work across fields. Like the randomization tests in the previous chapter, it approximates a reference distribution by resampling from the data; unlike randomization tests, it typically replaces an unknown population distribution with an estimated one. Today we cover the basic ideas behind the bootstrap and discuss under what (high-level) conditions we would expect it to work well in practice.

17.1 Confidence Sets

Let $X_i, i = 1, \ldots, n$ be an i.i.d. sample of observations with distribution $P \in \mathbf{P}$. The family $\mathbf{P}$ may be a parametric, nonparametric, or semiparametric family of distributions. We are interested in making inferences about some parameter

$$\theta(P) \in \Theta = \{\theta(P) : P \in \mathbf{P}\} \ .$$

To clarify the generality of the parameter $\theta(P)$, it is helpful to consider a few concrete examples:

- **Non-parametric Mean:** Let $X_i \in \mathbf{R}$ and consider the parameter

  $$\theta(P) = E_P[X_i] \ .$$

  This is perhaps the simplest possible case.

- **Slope Coefficient in a Linear Regression:** Suppose $X_i = (Y_i, W_i)$, where $Y_i \in \mathbf{R}$ is an outcome and $W_i \in \mathbf{R}^{k+1}$ is a vector of regressors. Let

  $$\theta(P) = E_P[W_i W_i']^{-1} E_P[W_i Y_i] \ ,$$

  which corresponds to the population coefficient from a linear projection of Y_i on W_i, assuming $E_P[W_i W_i']$ is invertible.

- **Average Treatment Effect (ATE):** Suppose $X_i = (Y_i, A_i)$, where $A_i \in \{0, 1\}$ is a binary treatment and Y_i is the observed outcome. Let $Y_i(1)$ and

$Y_i(0)$ denote potential outcomes, and assume random assignment. The parameter of interest is

$$\theta(P) = E_P[Y_i(1) - Y_i(0)] = E_P[Y_i \mid A_i = 1] - E_P[Y_i \mid A_i = 0] \ ,$$

where the second equality relies on the assumption that treatment is randomly assigned. (Here, we slightly abuse notation by using P to denote both the distribution of observed data and of potential outcomes; often the distribution over potential outcomes is denoted Q, and the mapping from Q to P depends on the treatment assignment mechanism.)

These examples show that $\theta(P)$ can refer to a simple moment, a parameter from a model that is vector-valued, or a causal estimand, depending on the context. More generally, $\theta(P)$ could be any function of P.

We are interested in constructing a confidence set for $\theta(P)$; that is, a random set, $C_n = C_n(X_1, \ldots, X_n)$ such that

$$P\{\theta(P) \in C_n\} \approx 1 - \alpha \ ,$$

at least for n sufficiently large.

The typical way of constructing such sets is based on approximating the distribution of a *root*, $R_n = R_n(X_1, \ldots, X_n, \theta(P))$. A root is simply any real-valued function depending on both the data $X_1, \ldots, X_n$ and the parameter of interest $\theta(P)$. The idea is that if the distribution of the root were known, then one could straightforwardly construct a confidence set for $\theta(P)$.

Example 17.1 Consider again the case where $\theta(P) = E_P[X_i]$, the mean of a real-valued random variable. A natural choice of root in this case is

$$R_n = \sqrt{n}(\bar{X}_n - \theta(P)) \ ,$$

where $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ is the sample mean. Notice that except for special circumstances, like when P is a normal distribution, the distribution of R_n is unknown. ■

Let $J_n(P)$ denote the sampling distribution of R_n, and define the corresponding cumulative distribution function as

$$J_n(x, P) = P\{R_n \leq x\} \ . \tag{17.1}$$

The notation is intended to emphasize that the distribution of the root depends on both the sample size n and the distribution P.

Using $J_n(x, P)$, we may choose a constant c such that

$$P\{R_n \leq c\} \approx 1 - \alpha \ .$$

Given such a c, the set

$$C_n = \{\theta \in \Theta : R_n(X_1, \ldots, X_n, \theta) \leq c\}$$

is a confidence set for $\theta(P)$ in the sense described above. In Example 17.1, this corresponds to

$$C_n = \left[\bar{X}_n - \frac{c}{\sqrt{n}}, \infty\right) \ .$$

If one prefers a two-sided interval, we may instead find c_1 and c_2 such that

$$P\{c_1 \le R_n \le c_2\} \approx 1 - \alpha \ ,$$

leading to

$$C_n = \{\theta \in \Theta : c_1 \le R_n(X_1, \ldots, X_n, \theta) \le c_2\} \ .$$

In Example 17.1, this yields the (almost) familiar construction

$$C_n = \left[\bar{X}_n - \frac{c_2}{\sqrt{n}}, \bar{X}_n - \frac{c_1}{\sqrt{n}}\right] \ .$$

The "almost" accounts for the fact that in the standard asymptotically normal case, the distribution is symmetric and so $c_1 = -c_2$. In general, the distribution of R_n need not be symmetric.

17.1.1 Pivots and Asymptotic Pivots

In some rare instances, $J_n(x, P)$ does not depend on P. In these instances, the root is said to be *pivotal* or a *pivot*. For example, if $\theta(P)$ is the mean of P and $\mathbf{P} = \{N(\theta, 1) : \theta \in \mathbf{R}\}$, then the root

$$R_n = \sqrt{n}(\bar{X}_n - \theta(P)) \tag{17.2}$$

is a pivot because $R_n \sim N(0, 1)$. In this case, we may construct confidence sets C_n with finite-sample validity; that is,

$$P\{\theta(P) \in C_n\} = 1 - \alpha$$

for all n and $P \in \mathbf{P}$.

Another example is the Kolmogorov-Smirnov statistic. If F is continuous and $\hat{F}_n$ is the empirical cdf, the distribution of

$$\sqrt{n} \sup_{x \in \mathbf{R}} |\hat{F}_n(x) - F(x)|$$

does not depend on the unknown F. Thus, even if its exact distribution is inconvenient, one can simulate the distribution under any convenient continuous distribution and use it to construct uniform confidence bands for F.

Sometimes, the root may not be pivotal in the sense described above, but it may be *asymptotically pivotal* or an *asymptotic pivot* in that $J_n(x, P)$ converges in distribution to a limit distribution $J(x, P)$ that does not depend

on P. For example, if $\theta(P)$ is the mean of P and $\mathbf{P}$ is the set of all distributions on $\mathbf{R}$ with a finite, nonzero variance, then

$$R_n = \frac{\sqrt{n}(\bar{X}_n - \theta(P))}{\hat{\sigma}_n} \tag{17.3}$$

is asymptotically pivotal because it converges in distribution to $J(x, P) = \Phi(x)$. In this case, we may construct confidence sets that are asymptotically valid in the sense that

$$\lim_{n\to\infty} P\{\theta(P) \in C_n\} = 1 - \alpha$$

for all $P \in \mathbf{P}$.

Example 17.2 (OLS Slope Coefficient) Suppose $\theta(P) = \beta_1$ is the slope coefficient from the linear regression model $Y_i = \beta_0 + \beta_1 W_i + U_i$, where $E_P[U_i W_i] = 0$. Let $\hat{\beta}_1$ denote the LS estimate and let $\hat{\sigma}_n^2$ be the usual heteroskedasticity-robust variance estimator for $\hat{\beta}_1$. Then the root

$$R_n = \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_n}$$

is not a pivot in finite samples because the distribution depends on P (through U_i and W_i), but under regularity conditions it converges in distribution to a standard normal:

$$R_n \xrightarrow{d} N(0, 1) \ .$$

Thus, R_n is an asymptotic pivot, and one can use it to construct asymptotically valid confidence intervals for β_1. ■

17.1.2 Asymptotic Approximations

Typically, the root will be neither a pivot nor an asymptotic pivot. The distribution of the root, $J_n(x, P)$, will typically depend on P, and, when it exists, the limit distribution of the root, $J(x, P)$, will, too. For example, if $\theta(P)$ is the mean of P and $\mathbf{P}$ is the set of all distributions on $\mathbf{R}$ with a finite, nonzero variance, then

$$R_n = \sqrt{n}(\bar{X}_n - \theta(P)) \tag{17.4}$$

converges in distribution to $J(x, P) = \Phi(x/\sigma(P))$. In this case, we can approximate this limit distribution with $\Phi(x/\hat{\sigma}_n)$, which will lead to confidence sets that are asymptotically valid in the sense described above.

Note that this third approach depends very heavily on the limit distribution $J(x, P)$ being both known and tractable. Even if it is known, the limit distribution may be difficult to work with (e.g., it could be the supremum of some complicated stochastic process with many nuisance parameters). Moreover, even if it is known and manageable, the method may be poor in finite-samples because it essentially relies on a double approximation: first, $J_n(x, P)$

is approximated by $J(x, P)$, then $J(x, P)$ is approximated in some way by estimating the unknown parameters of the limit distribution.

Example 17.3 Suppose $E[X_i] = \theta(P)$ and $\text{Var}(X_i) = \sigma_i^2 < \infty$ varies across i. The root

$$R_n = \sqrt{n}(\bar{X}_n - \theta(P))$$

converges in distribution to a normal distribution with variance $\lim n^{-1} \sum_{i=1}^n \sigma_i^2$, which depends on P and is unknown. ■

Example 17.4 Suppose we are interested in a scalar parameter $\theta(P)$ and know a priori that $\theta(P) \geq 0$. Let $\hat{\theta}_n$ be an estimator such that

$$\sqrt{n}(\hat{\theta}_n - \theta(P)) \xrightarrow{d} N(0, \sigma^2) \ .$$

Now consider the constrained estimator $\tilde{\theta}_n = \max\{\hat{\theta}_n, 0\}$, which incorporates the prior knowledge.

Then the asymptotic distribution of the constrained root

$$R_n = \sqrt{n}(\tilde{\theta}_n - \theta(P))$$

is no longer normal. When $\theta(P) > 0$, the constraint is inactive and $R_n \xrightarrow{d} Z := N(0, \sigma^2)$; but when $\theta(P) = 0$, the constraint binds and the limiting distribution becomes a folded normal. That is,

$$R_n \xrightarrow{d} \begin{cases} Z & \text{if } \theta(P) > 0 \\ \max\{Z, 0\} & \text{if } \theta(P) = 0 \end{cases} .$$

This distribution is not pivotal: it depends on σ^2 but, more importantly, on whether $\theta(P)$ is zero or not. It is not even smooth at the boundary. This setting arises in empirical work whenever inequality constraints are imposed for theoretical reasons (e.g., non-negative prices, demand slopes, entry games, auctions, etc). ■

17.2 The Bootstrap

The bootstrap is a fourth, more general approach to approximating $J_n(x, P)$. The idea is very simple: replace the unknown P with an estimate $\hat{P}_n$. Given $\hat{P}_n$, it is possible to compute (either analytically or using simulation to any desired degree of accuracy) $J_n(x, \hat{P}_n)$. In the case of i.i.d. data, a typical choice is the empirical distribution (though if $P = P(\psi)$ for some finite-dimensional parameter ψ, then one may also use $\hat{P}_n = P(\hat{\psi}_n)$ for some estimate $\hat{\psi}_n$ of ψ). The hope is that whenever $\hat{P}_n$ is "close" to P (which may be ensured, for

example, by the Glivenko-Cantelli Theorem), $J_n(x, \hat{P}_n)$ is "close" to $J_n(x, P)$. Essentially, this requires that $J_n(x, P)$, when viewed as a function of P, is continuous in an appropriate neighborhood of P. Often, this turns out to be true, but, unfortunately, it is not true in general.

17.2.1 The Nonparametric Mean

We will now consider the case where P is a distribution on $\mathbf{R}$ and $\theta(P)$ is the mean of P. We will consider first the root $R_n = \sqrt{n}(\bar{X}_n - \theta(P))$. Let $\hat{P}_n$ denote the empirical distribution of the $X_i, i = 1, \ldots, n$. Under what conditions is $J_n(x, \hat{P}_n)$ "close" to $J_n(x, P)$?

The sequence of distributions $\hat{P}_n$ is a random sequence, so it is more convenient to answer the question first for a nonrandom sequence P_n. The following theorem does exactly that.

Theorem 17.1 Let $\theta(P)$ be the mean of P and let $\mathbf{P}$ denote the set of all distributions on $\mathbf{R}$ with a finite, nonzero variance. Consider the root $R_n = \sqrt{n}(\bar{X}_n - \theta(P))$. Let $P_n, n \geq 1$ be a nonrandom sequence of distributions such that P_n converges in distribution to P, $\theta(P_n) \to \theta(P)$ and $\sigma^2(P_n) \to \sigma^2(P)$. Then,

(i) $J_n(x, P_n)$ converges in distribution to $J(x, P) = \Phi(x/\sigma(P))$.

(ii) $J_n^{-1}(1-\alpha, P_n) = \inf\{x \in \mathbf{R} : J_n(x, P_n) \geq 1-\alpha\}$ converges to

$$J^{-1}(1-\alpha, P) = z_{1-\alpha}\sigma(P) \ .$$

PROOF. (i) For each n, let $X_{i,n}, i = 1, \ldots, n$ be an i.i.d. sequence of random variables with distribution P_n. We must show that

$$\sqrt{n}(\bar{X}_{n,n} - \theta(P_n))$$

converges in distribution to $N(0, \sigma^2(P))$. To this end, let

$$Z_{n,i} = \frac{X_{n,i} - \theta(P_n)}{\sigma(P_n)}$$

and apply the Lindeberg-Feller central limit theorem. We must show that

$$\lim_{n\to\infty} E[Z_{n,i}^2 I\{|Z_{n,i}| > \epsilon\sqrt{n}\}] = 0 \ .$$

Let $\epsilon > 0$ be given. By the assumption that P_n converges in distribution to P and Slutsky's Theorem,

$$Z_{n,i} \xrightarrow{d} Z = \frac{X - \theta(P)}{\sigma(P)} \ ,$$

where X has distribution P. It follows that for any $\lambda > 0$ for which the distribution of $|Z|$ is continuous at λ, we have that

$$E[Z_{n,i}^2 I\{|Z_{n,i}| > \lambda\}] \to E[Z^2 I\{|Z| > \lambda\}] \ .$$

To prove this last claim, we need a couple of results:

1. Convergence of Moments [Lehmann and Romano, 2005, Example 11.2.14]: Suppose that Y_n and Y are real valued random variables and that $Y_n \xrightarrow{d} Y$. If the Y_n are uniformly bounded, then $E[Y_n] \to E[Y]$. (In general, convergence in distribution *does not* imply convergence of moments!)

2. Continuous Mapping Theorem [Lehmann and Romano, 2005, Theorem 11.2.13]: Suppose that $Y_n \xrightarrow{d} Y$. Let g be a measurable map from $\mathbf{R}$ to $\mathbf{R}$. Let C be the set of point in $\mathbf{R}$ for which g is continuous. If $P\{Y \in C\} = 1$, then $g(Y_n) \xrightarrow{d} g(Y)$.

We now use these two results. First, note that for any $\lambda > 0$ for which the distribution of $|Z|$ is continuous at λ, the continuous mapping theorem above implies that

$$g(|Z_{n,i}|) = Z_{n,i}^2 I\{|Z_{n,i}| \leq \lambda\} \xrightarrow{d} Z^2 I\{|Z| \leq \lambda\} = g(|Z|) \ . \tag{17.5}$$

Note that g is discontinuous at λ but that $P\{|Z| = \lambda\} = 0$, and so the result follows. Second, note that

$$E[Z_{n,i}^2 I\{|Z_{n,i}| > \lambda\}] = E[Z_{n,i}^2] - E[Z_{n,i}^2 I\{|Z_{n,i}| \leq \lambda\}] \ .$$

The first term on the right-hand side is always equal to one and also equal to $E[Z^2] = 1$. The second term is the expectation of

$$Z_{n,i}^2 I\{|Z_{n,i}| \leq \lambda\} \in [0, \lambda^2] \ ,$$

which is uniformly bounded. By (17.5) and the first result above,

$$E[Z_{n,i}^2 I\{|Z_{n,i}| \leq \lambda\}] \to E[Z^2 I\{|Z| \leq \lambda\}] \ .$$

We conclude that

$$\begin{aligned} E[Z_{n,i}^2 I\{|Z_{n,i}| > \lambda\}] &= E[Z^2] - E[Z_{n,i}^2 I\{|Z_{n,i}| \leq \lambda\}] \\ &\to E[Z^2] - E[Z^2 I\{|Z| \leq \lambda\}] \\ &= E[Z^2 I\{|Z| > \lambda\}] \ . \end{aligned}$$

As $\lambda \to \infty$, $E[Z^2 I\{|Z| > \lambda\}] \to 0$. To complete the proof, note that for any fixed $\lambda > 0$

$$E[Z_{n,i}^2 I\{|Z_{n,i}| > \epsilon\sqrt{n}\}] \leq E[Z_{n,i}^2 I\{|Z_{n,i}| > \lambda\}]$$

for n sufficiently large. Thus,

$$\sqrt{n}\bar{Z}_{n,n} \xrightarrow{d} N(0,1)$$

under P_n. The desired result now follows from Slutsky's Theorem and the fact that $\sigma(P_n) \to \sigma(P)$.

(ii) This follows from part (i) and Lemma 17.1 below applied to $F_n(x) = J_n(x, P_n)$ and $F(x) = J(x, P)$. ■

Lindeberg-Feller CLT

For each n, let $Z_{n,i}, i = 1, ..., n$ be i.i.d. with distribution P_n. Suppose $E_n[Z_{n,i}] = 0$ and $V_n[Z_{n,i}] = 1$. If for each $\epsilon > 0$

$$\lim_{n\to\infty} E_n[Z_{n,i}^2 I\{|Z_{n,i}| > \epsilon\sqrt{n}\}] = 0$$

then

$$\sqrt{n}\bar{Z}_{n,n} \xrightarrow{d} N(0,1) \quad \text{under} \quad P_n \ .$$

Lemma 17.1 Let $F_n, n \geq 1$ and F be nonrandom distribution functions on $\mathbf{R}$ such that F_n converges in distribution to F. Suppose F is continuous and strictly increasing at $F^{-1}(1-\alpha) = \inf\{x \in \mathbf{R} : F(x) \geq 1-\alpha\}$. Then, $F_n^{-1}(1-\alpha) = \inf\{x \in \mathbf{R} : F_n(x) \geq 1-\alpha\} \to F^{-1}(1-\alpha)$.

PROOF: Let $q = F^{-1}(1-\alpha)$. Fix $\delta > 0$ and choose ϵ so that $0 < \epsilon < \delta$ and F is continuous at $q-\epsilon$ and $q+\epsilon$. This is possible because F is continuous at q and therefore continuous in a neighborhood of q. Hence, $F_n(q-\epsilon) \to F(q-\epsilon) < 1-\alpha$ and $F_n(q+\epsilon) \to F(q+\epsilon) > 1-\alpha$, where the inequalities follow from the assumption that F is strictly increasing at q. For n sufficiently large, we thus have that $F_n(q-\epsilon) < 1-\alpha$ and $F_n(q+\epsilon) > 1-\alpha$. It follows that $q-\epsilon \leq F^{-1}(1-\alpha) \leq q+\epsilon$ for such n. ■

We are now ready to pass from the nonrandom sequence P_n to the random sequence $\hat{P}_n$.

Theorem 17.2 Let $\theta(P)$ be the mean of P and let $\mathbf{P}$ denote the set of all distributions on $\mathbf{R}$ with a finite, nonzero variance. Consider the root $R_n = \sqrt{n}(\bar{X}_n - \theta(P))$. Then,

(i) $J_n(x, \hat{P}_n)$ converges in distribution to $J(x, P) = \Phi(x/\sigma(P))$ a.s.

(ii) $J_n^{-1}(1-\alpha, \hat{P}_n)$ converges to $J^{-1}(1-\alpha, P) = z_{1-\alpha}\sigma(P)$ a.s.

PROOF: By the Glivenko-Cantelli Theorem,

$$\sup_{x\in\mathbf{R}} |\hat{P}_n((-\infty,x]) - P((-\infty,x])| \to 0 \quad a.s.$$

This implies that $\hat{P}_n$ converges in distribution to P a.s. Since $|x| \leq 1+x^2$ and that $\sigma^2(P) < \infty$, we have that $E[|X|] \leq 1 + E[X^2] < \infty$. Thus, we may apply the Strong Law of Large Numbers to conclude that $\theta(\hat{P}_n) = \bar{X}_n$ converges to $\theta(P)$ a.s. and $\sigma(\hat{P}_n)$ converges to $\sigma(P)$ a.s. Thus, w.p.1, $\hat{P}_n$ satisfies the assumptions of Theorem 17.1. The conclusions of the theorem now follow. ■

Remark 17.1 Similar results hold for the studentized root in (17.3) where $\hat{\sigma}_n$ is a consistent estimator of $\sigma(P)$. Using this root leads to the so-called Bootstrap-t, as the root is just the t-statistic. A key step in the proof of this result is to show that $\hat{\sigma}_n$ converges in probability to $\sigma(P)$ under an appropriate sequence of distributions; in the mean problem this follows from the same type of moment and uniform-integrability conditions used above. We skip the details in this class. However, the advantage of working with a studentized root like the one in (17.3) is that the limit distribution of R_n is pivotal, which affects the properties of the bootstrap approximation as discussed in the next section. ■

It now follows from Slutsky's Theorem that confidence sets of the form

$$C_n = \left\{\theta \in \mathbf{R} : R_n(X_1,\ldots,X_n,\theta) \leq J_n^{-1}(1-\alpha,\hat{P}_n)\right\} ,$$

which are known as *symmetric* confidence sets, or

$$C_n = \left\{\theta \in \mathbf{R} : J_n^{-1}\left(\frac{\alpha}{2},\hat{P}_n\right) \leq R_n(X_1,\ldots,X_n,\theta) \leq J_n^{-1}\left(1-\frac{\alpha}{2},\hat{P}_n\right)\right\} ,$$

which are known as *equi-tailed* confidence sets, satisfy

$$P\{\theta(P) \in C_n\} \to 1-\alpha \tag{17.6}$$

for all $P \in \mathbf{P}$.

In general, the consistency of the bootstrap is proved in the following two steps:

1. For some choice of metric (or pseudo-metric) d on the space of probability measures, it must be known that $d(P_n,P) \to 0$ implies that $J_n(P_n)$ converges weakly to $J(P)$. That is, the convergence of $J_n(P)$ to $J(P)$ must hold in a suitably locally uniform in P manner. After all, we are replacing P by $\hat{P}_n$ so $J_n(P)$ must be smooth in P. Note that in Theorem 17.1, the "metric" d that we used involved weak convergence together with convergence of first and second moments, see Remark 15.4.1 in Lehmann and Romano [2005] for details. However, other problems may require a different metric.

2. The estimator $\hat{P}_n$ must then be known to satisfy $d(\hat{P}_n,P) \to 0$ almost surely or in probability under P. This is what we proved in the proof of Theorem 17.2.

17.2.2 Asymptotic Refinements

Note that even a confidence set C_n based off of the asymptotic normality of either root would satisfy (17.6). It can be shown under certain conditions (that ensure the existence of so-called Edgeworth expansions of $J_n(x, P)$; see Section 15.5 of Lehmann and Romano (2005)) that one-sided confidence sets C_n based off of such an asymptotic approximation satisfy

$$P\{\theta(P) \in C_n\} - (1 - \alpha) = O(n^{-1/2}) \ . \tag{17.7}$$

One-sided confidence sets based off of the bootstrap and the root $R_n = \sqrt{n}(\bar{X}_n - \theta(P))$ also satisfy (17.7), though there is some evidence to suggest that it does a bit better in the size of $O(n^{-1/2})$ term. On the other hand, one-sided confidence sets based off the bootstrap-t, i.e., using the root

$$R_n = \frac{\sqrt{n}(\bar{X}_n - \theta(P))}{\hat{\sigma}_n}$$

as in Remark 17.1, satisfy

$$P\{\theta(P) \in C_n\} - (1 - \alpha) = O(n^{-1}) \ . \tag{17.8}$$

Thus, the one-sided coverage error of the bootstrap-t interval is $O(n^{-1})$ and is of smaller order than that provided by the normal approximation or the bootstrap based on a nonstudentized root. One-sided confidence sets that satisfy only (17.7) are said to be first-order accurate, whereas one-sided confidence sets that satisfy (17.8) are said to be second-order accurate. See Section 15.5 of Lehmann and Romano (2005) for further details.

A heuristic reason why the bootstrap based on the root (17.3) outperforms the bootstrap based on the root (17.2) is as follows. In the case of (17.2), the bootstrap is estimating a distribution that has mean 0 and unknown variance $\sigma^2(P)$. The main contribution to the estimation error is the implicit estimation of $\sigma^2(P)$ by $\sigma^2(\hat{P}_n)$. On the other hand, the root (17.3) has a distribution that is nearly independent of P since it is an asymptotic pivot.

The bootstrap may also provide a refinement in two-sided tests. For example, symmetric intervals based on the absolute value of the root in (17.3) are $O(n^{-2})$, versus the asymptotic approximation that is of order $O(n^{-1})$. Note that, by construction, such intervals are symmetric about $\hat{\theta}_n$.

17.2.3 Implementation of the Bootstrap

Outside certain exceptional cases, the bootstrap approximation $J_n(x, \hat{P}_n)$ cannot be calculated exactly, i.e., it is often not available in closed form. However, we can approximate this distribution to an arbitrary degree of accuracy by taking samples from $\hat{P}_n$, computing the root for each of these samples, and then using the empirical distribution of these roots as an approximation to $J_n(x, \hat{P}_n)$. The usual algorithm used to implement the bootstrap involves the following steps.

Step 1. Conditional on the data $(X_1, \ldots, X_n)$, draw B samples of size n from $\hat{P}_n$. Denote the jth sample by

$$(X_{1,j}^*, \ldots, X_{n,j}^*)$$

for $j = 1, \ldots, B$. When $\hat{P}_n$ is the empirical distribution, this amounts to resampling the original observations in $(X_1, \ldots, X_n)$ with replacement.

Step 2. For each bootstrap sample j, compute the root, i.e.,

$$R_{j,n}^* = R_n(X_{1,j}^*, \ldots, X_{n,j}^*, \hat{\theta}_n) \ .$$

Note that $\theta(\hat{P}_n) = \hat{\theta}_n$, so in the bootstrap distribution the parameter $\theta(P)$ becomes $\hat{\theta}_n$.

Step 3. Compute the empirical cdf of $(R_{1,n}^*, \ldots, R_{B,n}^*)$ as

$$L_n(x) = \frac{1}{B} \sum_{j=1}^{B} I\{R_{j,n}^* \leq x\} \ . \tag{17.9}$$

Step 4. Compute the desired function of $L_n(x)$, for example, a quantile,

$$L_n^{-1}(1-\alpha) = \inf\{x \in \mathbf{R} : L_n(x) \geq 1 - \alpha\} \ ,$$

for a given significance level α.

Remark 17.2 Sampling from $\hat{P}_n$ in Step 1 is easy even when $\hat{P}_n$ is the empirical distribution. In such case $\hat{P}_n$ is a discrete probability distribution that puts probability mass $\frac{1}{n}$ at each sample point $(X_1, \ldots, X_n)$, so sampling from $\hat{P}_n$ is equivalent to drawing observations (with probability $\frac{1}{n}$) from the observed data *with* replacement. In consequence, a bootstrap sample will likely have some ties and multiple values, which is generally not a problem. In parametric problems one would simply get a new sample of size n from $\hat{P}_n = P(\hat{\psi}_n)$. ■

Because B can be taken to be large (assuming enough computing power), the resulting approximation $L_n(x)$ can be made arbitrarily close to $J_n(x, \hat{P}_n)$. It then follows that the properties of tests and confidence sets based on $J_n^{-1}(1-\alpha, \hat{P}_n)$ and $L_n^{-1}(1-\alpha)$ are the same. In practice, values of B in the order of $1,000$ are frequently enough for the approximation to work well.

17.3 Concluding Remarks

The notes today follow the book by Lehmann and Romano (2005) Lehmann and Romano [2005] closely and notes kindly shared by Azeem Shaikh.

17.4 Problems

Problem 17.1 *Let $X_i, i = 1, \dots, n$ be i.i.d. Uniform$[0, \theta]$ with $\theta(P) = \theta > 0$, and let $\hat{\theta}_n = \max_{1 \le i \le n} X_i$. Consider the root*

$$R_n = n(\theta - \hat{\theta}_n) \ .$$

(a) Show that $R_n/\theta \xrightarrow{d} E$, where E is an Exponential random variable with mean 1. Hint: *compute $P\{R_n/\theta > t\} = P\{\hat{\theta}_n < \theta(1 - t/n)\}$ for $t \in (0, n)$ and take the limit.*

(b) Consider the nonparametric bootstrap: draw $X_1^*, \dots, X_n^*$ with replacement from $(X_1, \dots, X_n)$, let $\hat{\theta}_n^* = \max_i X_i^*$, and define the bootstrap root $R_n^* = n(\hat{\theta}_n - \hat{\theta}_n^*)$. Show that, conditional on the data,

$$P\{\hat{\theta}_n^* = \hat{\theta}_n\} = 1 - \left(1 - \frac{1}{n}\right)^n \to 1 - e^{-1} \ .$$

(c) Conclude that the bootstrap distribution of R_n^* places mass approximately $1 - e^{-1} \approx 0.632$ at the point 0 for all large n, and therefore cannot converge to the continuous limit $J(x, P)$ of part (a). Explain how this relates to the requirement, stated in the chapter, that $J_n(x, P)$ be continuous in P in an appropriate neighborhood for the bootstrap to be consistent.

Problem 17.2 *For each statement, say whether it is true or false and justify your answer in one or two sentences. Assume throughout the i.i.d. nonparametric-mean setting of the chapter, with $\theta(P) = E_P[X_i]$ and a finite, nonzero variance.*

(a) If a root is asymptotically pivotal, then the bootstrap is automatically consistent for its sampling distribution.

(b) For one-sided confidence intervals, the bootstrap-t based on the studentized root $\sqrt{n}(\bar{X}_n - \theta(P))/\hat{\sigma}_n$ has coverage error $O(n^{-1})$, an improvement over the $O(n^{-1/2})$ error of the normal approximation.

(c) The bootstrap based on the non-studentized root $\sqrt{n}(\bar{X}_n - \theta(P))$ attains the same order of one-sided coverage error as the bootstrap-t.

(d) Symmetric two-sided intervals based on the absolute studentized root have coverage error $O(n^{-2})$.

For (c), identify which quantity the non-studentized bootstrap must implicitly estimate, and explain why that estimation is the dominant source of error.

Bibliography

B. Hansen. *Econometrics*. Princeton University Press, 2022.

E. Lehmann and J. P. Romano. *Testing Statistical Hypotheses*. Springer, New York, 3rd edition, 2005.

D. N. Politis, J. P. Romano, and M. Wolf. *Subsampling*. Springer, New York, 1999.

18

Inference with Clustered Data

Let (Y, X, U) be a random vector where Y and U take values in $\mathbf{R}$ and X takes values in $\mathbf{R}^{k+1}$. Assume further that the first component of X is a constant equal to one, i.e., $X = (X_0, X_1, \ldots, X_k)'$ with $X_0 = 1$. Let $\beta \in \mathbf{R}^{k+1}$ be such that

$$Y = X'\beta + U \ . \tag{18.1}$$

Suppose we observe a sample of size n in a context where the variables are grouped into q mutually independent known *clusters*, indexed by $j = 1, \ldots, q$. The clusters can be due to sampling scheme or by the researcher knowing the correlation structure of the observed sample. We will take a conservative stance within a cluster, which means we do not assume any ordering inside a cluster because there may not be one.

In this class we discuss two approaches for inference on β. The first approach is based on asymptotic normality and a consistent estimator of the asymptotic variance, in an asymptotic framework where the number of clusters grows to infinity. To this end, we develop appropriate Law of Large Numbers (LLN) and Central Limit Theorem (CLT) for dependent processes that mimic the clustered data. We then use these results to show that the Cluster Covariance Estimator (CCE) for the asymptotic variance of β is consistent and leads to an asymptotically valid test. The second approach is based on the wild bootstrap, and is particularly appealing when the number of clusters is not sufficiently large. Indeed, we present the properties of this test in an asymptotic framework where the number of observations *within* the clusters is large, but the actual number of clusters is fixed.

18.1 Setup and Notation

We use two indices as we did when we covered panel data. Denote by $j = 1, \ldots, q$ the index for the clusters. Denote by $i = 1, \ldots, n_j$ the units within cluster j. For instance, a cluster could be a family, a school, an industry, or a city. A unit could be a family member, a student, a firm, or a citizen, respectively.

For each cluster j, let us denote by $X_j = (X_{1,j}, \ldots, X_{n_j,j})' \in \mathbf{R}^{n_j \times (k+1)}$ the matrix of stacked observations. Define $Y_j \in \mathbf{R}^{n_j}$ and $U_j \in \mathbf{R}^{n_j}$ in a similar

way. Note that (18.1) implies

$$Y_j = X_j\beta + U_j, \quad j = 1, \ldots, q, \quad \text{where} \quad E[X_j'U_j] = 0 \ .$$

We assume that (X_j, U_j, Y_j) are independent across clusters but remain agnostic about the dependence within clusters.

18.2 Law of Large Numbers

We temporarily focus on the sample mean of $X_{i,j}$ since LS estimator is a function of sample means:

$$\bar{X}_n = \frac{1}{n}\sum_{j=1}^{q} X_j'\mathbf{1}_j = \frac{1}{n}\sum_{j=1}^{q}\sum_{i=1}^{n_j} X_{i,j} \ ,$$

where $\mathbf{1}_j$ is an n_j-dimensional vector of ones. Hansen and Lee [2019] prove the following result.

Theorem 18.1 *Suppose that as* $n \to \infty$,

$$\max_{j \le q} \frac{n_j}{n} \to 0 \tag{18.2}$$

and that

$$\lim_{M\to\infty} \sup_{i,j} E\left[||X_{i,j}||I\{||X_{i,j}|| > M\}\right] = 0 \ . \tag{18.3}$$

Then, as $n \to \infty$

$$||\bar{X}_n - E\left[\bar{X}_n\right]|| \xrightarrow{P} 0 \ .$$

A few comments are in order. The condition in (18.3) states that $X_{i,j}$ is uniformly integrable (similar to other standard conditions). Assumption (18.2) states that each cluster size n_j is asymptotically negligible. This condition automatically holds when n_j is fixed as $q \to \infty$ (say, the original framework used to study clustered data). In particular, notice that it implies $q \to \infty$, so there is no need to explicitly write this condition. Assumption (18.2) allows for considerable heterogeneity in cluster sizes and it allows the cluster sizes to grow with sample size, so long as the growth is not proportional. It is necessary for parameter estimation consistency while allowing arbitrary within-cluster dependence. Otherwise a single cluster could dominate the sample average.

As a convenient example, consider the case where $n_j = n^a$ for $0 \le a < 1$, which leads to

$$n = \sum_{j=1}^{q} n_j = \sum_{j=1}^{q} n^a = qn^a \implies q = n^{1-a} \ .$$

18.3 Central Limit Theorem

Under i.i.d. sampling the standard deviation of the sample mean, $\bar{X}_n$, is of order $O(n^{-1/2})$, so $\sqrt{n}$ is the natural scaling to obtain the central limit theorem (CLT). However, clustering often alters the rate of convergence and may lead to rates of convergence of order $O(q^{-1/2})$ or slower. See Section 18.7 for a variety of examples, including a heterogeneous-size design where the rate is slower than both $n^{-1/2}$ and $q^{-1/2}$. In settings where the rate of convergence may not be easily pinned down, it becomes essential to standardize the sample mean (or the estimators under consideration) by the actual square root of the variance.

Define the variance-covariance matrix of $\sqrt{n}\bar{X}_n$ by

$$\begin{aligned}\Omega_n &:= E\left[n\left(\bar{X}_n - E[\bar{X}_n]\right)\left(\bar{X}_n - E[\bar{X}_n]\right)'\right] \\ &= \frac{1}{n}\sum_{j=1}^{q} E\left[\left(X_j'\mathbf{1}_j - E[X_j'\mathbf{1}_j]\right)\left(X_j'\mathbf{1}_j - E[X_j'\mathbf{1}_j]\right)'\right] ,\end{aligned}$$

where $\mathbf{1}_j$ is an n_j-dimensional vector of ones.

The next theorem presents a central limit theorem for the sample mean considering the right scaling, $\Omega_n^{-1/2}\sqrt{n}$. By construction,

$$\Omega_n^{-1/2}\sqrt{n}(\bar{X}_n - E[\bar{X}_n])$$

is a random variable with zero mean and a covariance matrix equal to the identity matrix. In the statement of the theorem, we use $\lambda_n = \lambda_{\min}(\Omega_n)$ to denote the minimum eigenvalue of Ω_n and $\mathbb{I}_{k+1}$ to denote the identity matrix of dimension $k+1$.

Theorem 18.2 *Suppose that for some* $2 \leq r < +\infty$,

$$\lim_{M\to\infty} \sup_{i,j} E\left[||X_{i,j}||^r I\{||X_{i,j}|| > M\}\right] = 0 , \tag{18.4}$$

and

$$\frac{\left(\sum_{j=1}^{q} n_j^r\right)^{2/r}}{n} \leq C < \infty , \tag{18.5}$$

for some positive $C > 0$. *Assume further that as* $n \to \infty$,

$$\max_{j\leq q} \frac{n_j^2}{n} \to 0 \tag{18.6}$$

and

$$\lambda_n \geq \lambda > 0 , \tag{18.7}$$

for some positive $\lambda > 0$. *Then, as* $n \to \infty$

$$\Omega_n^{-1/2}\sqrt{n}\left(\bar{X}_n - E[\bar{X}_n]\right) \xrightarrow{d} N(0, \mathbb{I}_{k+1}) . \tag{18.8}$$

Now let us discuss the conditions under which this theorem holds. Assumption (18.4) states that $||X_{i,j}||^r$ is uniformly integrable. When $r = 2$, this condition is similar to the Lindeberg condition for the CLT under independent heterogeneous sampling. Assumption (18.5) involves a trade-off between the cluster sizes and the number of moments r. It is less restrictive for large r, and more restrictive for small r. Note that as $r \to \infty$, we can conclude $\max_{j \leq q} n_j^2/n = O(1)$, which is implied by (18.6).

Assumption (18.6) allows for growing and heterogeneous cluster sizes. It allows clusters to grow uniformly at the rate $n_j = n^a$ for any $0 \leq a \leq (r-2)/(2r-2)$. Note that this requires the cluster sizes to be bounded if $r = 2$. It also allows for only a small number of clusters to grow. For example, $n_j = \bar{n}$ (bounded clusters) for $q-k$ clusters and $n_j = q^{a/2}$ for k clusters, with k fixed. In this case the assumption holds for any $a < 1$ and $r = 2$.

Finally, Assumption (18.7) specifies that $\text{Var}[\sqrt{n}c'\bar{X}_n]$ does not vanish for any vector $c \neq 0$, since the condition implies that the minimum eigenvalue of the variance-covariance matrix is positive.

18.3.1 Cluster Covariance Estimation

Theorem 18.2 implies that the right scaling for $\bar{X}_n$ is $\Omega_n^{-1/2}\sqrt{n}$. However, since Ω_n is typically unknown, this scaling is not particularly useful in practice. To deal with this, we now consider the estimation of Ω_n and see if by replacing it with a consistent estimator we retain the same level of convergence. Suppose that $E[X_j'\mathbf{1}_j] = 0$ for all $j = 1, \ldots, q$. This centering assumption is only an exposition device: in general one can apply the same argument to the centered cluster sums $X_j'\mathbf{1}_j - E[X_j'\mathbf{1}_j]$. Under this simplification,

$$\Omega_n = \frac{1}{n}\sum_{j=1}^{q} E\left[(X_j'\mathbf{1}_j - E[X_j'\mathbf{1}_j])(X_j'\mathbf{1}_j - E[X_j'\mathbf{1}_j])'\right]$$

is equal to

$$\frac{1}{n}\sum_{j=1}^{q} E[X_j'\mathbf{1}_j\mathbf{1}_j'X_j] \ .$$

This has the following natural estimator

$$\widehat{\Omega}_n = \frac{1}{n}\sum_{j=1}^{q} X_j'\mathbf{1}_j\mathbf{1}_j'X_j = \frac{1}{n}\sum_{j=1}^{q}\left(\sum_{i=1}^{n_j} X_{i,j}\right)\left(\sum_{i=1}^{n_j} X_{i,j}\right)' \ .$$

It is worth to mention that this estimator allows for arbitrary within-cluster dependence patterns. Also, it allows for heterogeneity since $E[X_j'\mathbf{1}_j\mathbf{1}_j'X_j]$ can vary across j. The following theorem presents the relevance of this estimator to obtain the asymptotic normality of the sample mean $\bar{X}_n$ after the right scaling. The proof of the theorem is in Hansen and Lee [2019].

Theorem 18.3 (Consistency of CCE) *Under the same assumptions of Theorem 18.2 and assuming that* $E[X_j' \mathbf{1}_j] = 0$*, we obtain as* $n \to \infty$ *that*

$$||\widehat{\Omega}_n - \Omega_n|| \xrightarrow{P} 0$$

and

$$\widehat{\Omega}_n^{-1/2} \sqrt{n} \bar{X}_n \xrightarrow{d} N(0, \mathbb{I}_{k+1}) \ .$$

18.4 Inference for Linear Regression: CCE Approach

Let us recall our initial setup. For each cluster j, let us denote by $X_j = (X_{1,j}, \ldots, X_{n_j,j})' \in \mathbf{R}^{n_j \times (k+1)}$ the matrix of stacked observations. Define $Y_j \in \mathbf{R}^{n_j}$ and $U_j \in \mathbf{R}^{n_j}$ in a similar way. Using this notation, we have

$$Y_j = X_j \beta + U_j, \quad j = 1, \ldots, q, \quad \text{where} \quad E[X_j' U_j] = 0 \ ,$$

and we assume that (Y_j, X_j) are independent across clusters but remain agnostic about the dependence within clusters.

The least square (LS) estimator of β is given by

$$\hat{\beta}_n = \left(\frac{1}{n} \sum_{j=1}^{q} X_j' X_j \right)^{-1} \frac{1}{n} \sum_{j=1}^{q} X_j' Y_j \ .$$

Using this expression and the model for Y_j, we can derive the following

$$\sqrt{n}(\hat{\beta}_n - \beta) = \left(\frac{1}{n} \sum_{j=1}^{q} X_j' X_j \right)^{-1} \frac{1}{\sqrt{n}} \sum_{j=1}^{q} X_j' U_j \ .$$

In order to discuss the consistency and the asymptotic normality properties of the LS estimator, consider the following notation:

$$\Sigma_n = \frac{1}{n} \sum_{j=1}^{q} E[X_j' X_j] \quad \text{and} \quad \Omega_n = \frac{1}{n} \sum_{j=1}^{q} E[X_j' U_j U_j' X_j] \ .$$

Consistency of LS: If the condition (18.2) in Theorem 18.1 holds, Σ_n has full rank, $\lambda_{min}(\Sigma_n) \geq C > 0$, and the uniform integrability condition in (18.3) holds for $X_{i,j} X_{i,j}'$ and $X_{i,j}' U_{i,j}$, then

$$\hat{\beta}_n \xrightarrow{P} \beta \quad \text{as} \quad n \to \infty \ .$$

Asymptotic Normality of LS: To properly normalize $\sqrt{n}(\hat{\beta}_n - \beta)$ we define

$$\mathbb{V}_n = \Sigma_n^{-1} \Omega_n \Sigma_n^{-1}$$

as the rate of convergence may not be $\sqrt{n}$. Using this notation, we assume that the conditions in Theorem 18.2 hold for some r, Σ_n has full rank, $\lambda_{\min}(\Sigma_n) \geq C > 0$, $\lambda_{\min}(\Omega_n) \geq C > 0$, and the uniform integrability condition in (18.4) holds for $X_{i,j}X_{i,j}'$ and $X_{i,j}'U_{i,j}$. It follows that as $n \to \infty$:

$$\mathbb{V}_n^{-1/2}\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} N(0, \mathbb{I}_{k+1}) \ .$$

18.4.1 Cluster Covariance Estimator

The final piece is a consistent estimator for $\mathbb{V}_n$ that would allow us to do inference on the LS estimator.

Definition 18.1 (Cluster Covariance Estimator: CCE) *Denote by $\widehat{\mathbb{V}}_n$ the CCE estimator, that is*

$$\widehat{\mathbb{V}}_n = \left(\frac{1}{n}\sum_{j=1}^{q} X_j'X_j\right)^{-1} \frac{1}{n}\sum_{j=1}^{q} X_j'\hat{U}_j\hat{U}_j'X_j \left(\frac{1}{n}\sum_{j=1}^{q} X_j'X_j\right)^{-1} ,$$

where $\hat{U}_j = Y_j - X_j\hat{\beta}_n$ are the LS residuals.

Under the same conditions listed for the asymptotic normality of the LS estimator, we can conclude that as $n \to \infty$,

$$||\widehat{\mathbb{V}}_n - \mathbb{V}_n|| \xrightarrow{P} 0 \quad \text{and} \quad \widehat{\mathbb{V}}_n^{-1/2}\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} N(0, \mathbb{I}_{k+1}) \ .$$

Note that in the special case with $n_j = 1$ for all $j = 1, \ldots, q$, this estimator becomes the HC estimator. It is worth mentioning that `Stata` uses a multiplicative adjustment to reduce the bias,

$$\widehat{\mathbb{V}}_{\text{stata}} = \frac{n-1}{n-k-1}\frac{q}{q-1}\widehat{\mathbb{V}}_n \ .$$

18.4.2 Inference

For $s \in \{0, 1, \ldots, k\}$, let β_s be the s-th element of β and let $\widehat{\mathbb{V}}_{n,s}$ be the $(s+1)$-th diagonal element of $\widehat{\mathbb{V}}_n$. Using this notation, consider testing

$$H_0 : \beta_s = c \quad \text{versus} \quad H_1 : \beta_s \neq c$$

at level α. Using the results we just derived, it follows that under the null hypothesis, the t-statistic is asymptotically standard normal,

$$t_{\text{stat}} = \frac{\sqrt{n}(\hat{\beta}_{n,s} - c)}{\sqrt{\widehat{\mathbb{V}}_{n,s}}} \xrightarrow{d} N(0,1) \quad \text{as} \quad n \to \infty \ .$$

This implies that the test that rejects H_0 when $|t_{\text{stat}}| > z_{1-\alpha/2}$ is consistent in levels, where $z_{1-\alpha/2}$ is a critical value defined by the $(1-\alpha/2)$-quantile of the standard normal distribution.

18.4.3 The Problem with Few Clusters

The asymptotic derivations we have derived so far assume that the total number of observations, n, is large. This, in combination with the conditions required for the LLN and the CLT immediately imply that the number of clusters, q, is also assumed to be large. In empirical settings, it is however not unusual to find instances where the actual number of clusters is rather small (say, few regions, few schools, few states, etc). It turns out that the performance of t-test with CCE can be quite poor when q is small, even if n is large. We illustrate this with a simple simulation.

There are finite-sample adjustments that people use for small q in practice. For instance, Bell and McCaffrey [2002] propose a bias-reduction modification analogous to HC2 and a t critical value with a degrees-of-freedom adjustment, following the same intuition as in Behrens-Fisher problems. The $q-1$ rows in Table 18.1 should be read in this spirit: they replace the normal critical value with a more conservative critical value when the effective number of independent pieces of information is closer to the number of clusters.

Table 18.1 reports simulation results for the following five designs. The model in all designs is

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + U_i \\ X_i &= V_{C_i} + W_i \\ U_i &= \nu_{C_i} + \eta_i \ , \end{aligned}$$

where C_i denotes the cluster of i and all variables are $N(0,1)$. Note that the components V_{C_i} and ν_{C_i} that are common among all units within a cluster lead to cluster-level dependence. Also, in all designs $\beta_0 = \beta_1 = 0$. In the first design there are $q = 10$ clusters, with $n_j = 30$ units in each cluster. In the second design $q = 5$ with $n_j = 30$. In the third design there are $q = 10$ clusters, half with $n_j = 10$ and half with $n_j = 50$. The fourth design has heteroskedasticity, with $\eta_i | X_i \sim N(0, 0.9X_i^2)$, and the fifth design, the covariate is fixed within the clusters: $W_i = 0$ and $V_{C_i} \sim N(0,2)$. The last two designs have $q = 10$ clusters with $n_j = 30$.

The following table reports the coverage probability of the confidence intervals, using three cluster covariance estimators: (a) the standard CCE, (b) the one used in Stata, and (c) a bias-reduction modification analogous to that of HC2 that we denote by $\hat{V}_{\text{cce2}}$.

When q is small, $\hat{\mathbb{V}}_{cce2}$ (more so $\hat{\mathbb{V}}_n$ and $\hat{\mathbb{V}}_{\text{stata}}$) typically leads to confidence sets that under-cover. Using a critical value from a t-distribution with $q-1$ degrees of freedom (dof) instead of the usual standard normal critical value (dof=∞) reduces the amount of under-coverage but it still leads to coverage probabilities that are below the nominal level (95%). A recent contribution by Hansen [2026] provides a theoretical foundation for using jackknife cluster standard errors in this setting. In particular, Hansen shows that the conventional cluster-robust variance estimator can have arbitrarily large downward bias, while a carefully defined jackknife cluster variance estimator is never

TABLE 18.1: Design in Imbens and Kolesar/CGM: $1 - \alpha = 95\%$

	dof	I	II	III	IV	V
$\hat{V}_n$	∞	84.7	73.9	79.6	85.7	81.7
	q-1	89.5	86.9	85.2	90.2	86.4
$\hat{V}_{\text{stata}}$	∞	86.7	78.8	81.9	87.6	83.6
	q-1	91.1	90.3	87.2	91.8	88.1
$\hat{V}_{\text{cce2}}$	∞	89.2	84.7	87.2	89.1	87.7
	q-1	93.0	93.3	91.3	92.8	91.4

downward biased for the finite-sample variance under broad conditions. He also proposes pairing this jackknife standard error with a Satterthwaite-type degrees-of-freedom adjustment. These results reinforce the main lesson from the simulation above: with few or influential clusters, conventional clustered standard errors and normal critical values can be misleading. This has led the literature to explore alternatives to the standard asymptotic approximation when the number of clusters is small.

18.5 The Wild Bootstrap

A better approach to inference when q is small is to use methods that can be theoretically justified in an asymptotic framework where $n_j \to \infty$ for each cluster but q, the number of cluster, remains bounded (i.e., fixed). These include, for example, the approach based on "approximate" randomization tests originally proposed by Canay et al. [2017] and further studied by Cai et al. [2023]. This approach has the advantage of not imposing assumptions on the degree of heterogeneity across clusters. In practice, perhaps the most popular approach is the wild bootstrap test proposed by Cameron et al. [2008]. This method was proposed on the grounds of superior performance in simulations but its formal properties were unknown until the work by Canay et al. [2021], who provided conditions under which this test is valid in an asymptotic framework where the number of clusters q is fixed.

18.5.1 The test

Under the framework in Section 18.1, assume that in addition to $X_{i,j} \in \mathbf{R}^{d_x}$, we have regressors $W_{i,j} \in \mathbf{R}^{d_w}$, such as additional unit-level characteristics or cluster-level fixed effects. Here $X_{i,j}$ denotes only the regressors of interest, while the constant and other controls are collected in $W_{i,j}$. The model can be

then written as

$$Y_{i,j} = X'_{i,j}\beta + W'_{i,j}\gamma + U_{i,j} \ ,$$

where $\beta \in \mathbf{R}^{d_x}$ is the parameter of interest and $\gamma \in \mathbf{R}^{d_w}$ is a nuisance parameter in the sense that it is associated with the control variable $W_{i,j}$.

Suppose we aim to test

$$H_0 : c'\beta = \lambda \quad \text{versus} \quad H_1 : c'\beta \neq \lambda,$$

for given values of $c \in \mathbf{R}^{d_x}$ and $\lambda \in \mathbf{R}$ at level α. The test statistic is

$$T_n := \left| \sqrt{n} \left(c'\hat{\beta}_n - \lambda \right) \right| ,$$

where $\hat{\beta}_n$ and $\hat{\gamma}_n$ are the OLS estimators of β and γ. The critical value is constructed in the following way.

Step 1 Compute $\hat{\beta}^{\mathrm{r}}_n$ and $\hat{\gamma}^{\mathrm{r}}_n$, the restricted least squares estimators of β and γ obtained under the constraint that $c'\beta = \lambda$. Note that $c'\hat{\beta}^{\mathrm{r}}_n = \lambda$ by construction.

Step 2 Let $\mathbf{G} = \{-1, 1\}^q$ and for any $g = (g_1, \ldots, g_q) \in \mathbf{G}$, define

$$Y^*_{i,j}(g) \equiv X'_{i,j}\hat{\beta}^{\mathrm{r}}_n + W'_{i,j}\hat{\gamma}^{\mathrm{r}}_n + g_j \hat{U}^{\mathrm{r}}_{i,j},$$

where

$$\hat{U}^{\mathrm{r}}_{i,j} = Y_{i,j} - X'_{i,j}\hat{\beta}^{\mathrm{r}}_n - W'_{i,j}\hat{\gamma}^{\mathrm{r}}_n \ .$$

For each $g \in \mathbf{G}$, compute $\hat{\beta}^*_n(g)$ and $\hat{\gamma}^*_n(g)$, the OLS estimators of β and γ obtained by using $Y^*_{i,j}(g)$ in place of $Y_{i,j}$ and the same regressors $\left(X'_{i,j}, W'_{i,j}\right)'$.

Step 3 Compute the $1-\alpha$ quantile of $\left\{ \left| \sqrt{n} c' \left(\hat{\beta}^*_n(g) - \hat{\beta}^{\mathrm{r}}_n \right) \right| : g \in \mathbf{G} \right\}$, denoted by $\hat{c}_n(1-\alpha)$.

The test is then defined as follows,

$$\phi^{\mathrm{wild}}_n := I\left\{ T_n > \hat{c}_n(1-\alpha) \right\}.$$

The bootstrap sample is constructed in Step 2 in a similar way as in the sign-change randomization tests covered earlier in the resampling part. Specifically, we consider the group of sign change transformations $\mathbf{G}$ applied to the residuals of the restricted OLS. In fact, the validity of this test is proved by showing the limiting rejection probability of ϕ^{wild}_n equals that of a level-α randomization test.

18.5.2 Asymptotic validity of the test

It is worth noting that now the asymptotic framework is to fix q and let $n_j \to \infty$ for all $j = 1, \ldots, q$. Two sets of assumptions are needed for the asymptotic analysis of the test. We will only briefly discuss the implications of these assumptions and the details are in Canay et al. [2021].

The first set of assumptions (Assumption LS) imposes sufficient conditions to ensure that the OLS estimators and the restricted OLS estimators of β and γ are well behaved. It is satisfied, for example, whenever the within-cluster dependence is sufficiently weak to permit application of suitable laws of large numbers and central limit theorems. We omit the details here.

The second set of assumptions (Assumption H) require a certain degree of *homogeneity* across clusters (in terms of their size, but also in terms of the distributions of the covariates). Conceptually, they require that the distribution of $\left(X'_{i,j}, W'_{i,j}\right)'$ and $\left(X'_{i,j'}, W'_{i,j'}\right)'$ for $j \neq j'$ are not too different in a certain sense that we discuss below.

Under the aforementioned assumptions, Canay et al. [2021] proved the following result

Theorem 18.4 *If Assumptions LS and H hold and $c'\beta = \lambda$, then*

$$\alpha - \frac{1}{2^{q-1}} \leq \lim_{n\to\infty} E[\phi_n^{\text{wild}}] \leq \alpha \ .$$

Note that the theorem not only shows that the limiting rejection probability is bounded above by α, but also that it is bounded below by $\alpha - \frac{1}{2^{q-1}}$, implying that the test cannot be "too" conservative.

The main requirement for the wild bootstrap to work well with a few number of clusters is Assumption H. In order to discuss this assumption, let

$$\tilde{X}_{i,j} := X_{i,j} - \hat{\Pi}'_n W_{i,j}$$

where $\hat{\Pi}_n$ is the matrix of coefficient obtained from linearly projecting $X_{i,j}$ on $W_{i,j}$, i.e.,

$$\sum_{j=1}^{q} \sum_{i=1}^{n_j} \left(X_{i,j} - \hat{\Pi}'_n W_{i,j}\right) W'_{i,j} = 0 \ .$$

Using this notation, the requirement is that for each $j = 1, \ldots, q$,

$$\frac{1}{n_j} \sum_{i=1}^{n_j} \tilde{X}_{i,j} \tilde{X}'_{i,j} \xrightarrow{P} a_j \Omega_{\tilde{X}} \ ,$$

where $a_j > 0$ and $\Omega_{\tilde{X}}$ positive definite. Intuitively, the variance-covariance matrix of $\tilde{X}_{i,j}$ has to be proportional across clusters; a strong restriction on the type of heterogeneity allowed across clusters.

What are the main implications for empirical practice? Theorem 18.4

shows that wild bootstrap-based tests can be valid and perform well when the number of clusters is small, but there are two caveats that should be kept in mind. First, formal results only hold for a specific variant of the wild bootstrap-based test proposed in Cameron et al. [2008]. Practitioners should, in particular, use the weights specified in Step 2 as the result in Theorem 18.4 only applies to these weights. Second, the validity of the wild bootstrap-based tests rely on certain homogeneity assumptions on the distribution of covariates across clusters. These homogeneity requirements can sometimes be weakened by including cluster-level fixed effects (so this is generally recommended), but may fail in many settings. Whenever the number of clusters is small and the homogeneity assumptions are implausible, however, there are other inference procedures available that do not rely on such homogeneity conditions, such as the exact t-approach (See Ibragimov and Müller [2010]) and approximate randomization tests (See Canay et al. [2017]).

18.6 When and At What Level to Cluster?

The choice of whether and how to cluster standard errors has important implications for inference in empirical research. So far we have studied the conventional econometric framework for clustering, which focuses on adjusting for correlation induced by unobserved cluster-level components in the sampling process for the outcome variable. However, this framework leaves some key questions unanswered: Why do we adjust standard errors for clustering in some ways but not others, for example, by state but not by gender, and in observational studies but not in completely randomized experiments? In what settings does the choice of whether and how to cluster make a difference?

Abadie et al. [2023] address these questions for clustered inference on average treatment effects. They propose a new framework that incorporates both a sampling component (how units are drawn into the sample) and a design component (how treatment is assigned across units and clusters). The key message is that the need to cluster standard errors, and the level at which to cluster, is primarily driven by these two components rather than by unobserved cluster-level errors in the outcome model.

Specifically, the authors argue that clustering is required when:

1. There is variation in treatment assignment probabilities across clusters (the "design" component). Even with random sampling of units, if treatment is assigned at the cluster level, Abadie et al. [2023] argue that standard errors should be clustered.

2. A non-negligible fraction of the total clusters in the population are sampled (the "sampling" component). If all or nearly all clusters are sampled, Abadie et al. [2023] argue that standard errors should be clustered. This

second consideration of course relies on a finite-population framework, as opposed to a super-population framework.

Beyond the average treatment effect in the type of setting considered by Abadie et al. [2023], the question of when and at what level to cluster depends on the characteristics of the DGP under consideration. Perhaps the most important feature is to understand the underlying trade-off behind this choice: the less we cluster (say, in the limit case by assuming the observations are i.i.d.), the more we demand from the data in terms of information (i.e., independence). As a result, the effective number of observations is "large". On the other hand, the more we cluster (say, at the state level in a data set with individuals, in cities, in counties, in states), the more agnostic we are about the intrinsic dependence in the data. As a result, the effective number of observations becomes "small" (i.e., the number of states).

As a practical rule, clustering is most compelling when treatment assignment varies at the cluster level, when the sampling process draws a non-negligible share of population clusters, or when the maintained sampling model allows arbitrary dependence within groups. The level of clustering should be chosen to match the source of dependence one is unwilling to rule out; clustering more coarsely is more robust, but it also reduces the effective sample size.

18.7 Supplement: rates of convergence

Under i.i.d. sampling the rate of convergence of the sample mean is $n^{-1/2}$. That is

$$\sqrt{n}\left(\bar{X}_n - E\left[\bar{X}_n\right]\right) \xrightarrow{d} N(0, V) \ .$$

In the case of clustering data, the rate of convergence may or may not be affected. We will see in the example below that often is affected. For instance, if the dependence within the cluster is strong, the rate of convergence is determined by the number of clusters: $q^{-1/2}$. If the dependence within clusters is weak —in a precise sense that we illustrate later in the examples—, the rate of convergence is $n^{-1/2}$. These examples are based on Hansen and Lee [2019].

However, if the dependence is in between weak and strong, the rate of convergence can be in between or even slower than the rates mentioned above. This result was shown by Hansen and Lee [2019] and we will illustrate it also in the following examples.

To analyse the convergence rate, we can compute the standard deviation

of the sample mean. That is

$$\begin{aligned}\mathrm{sd}(\bar{X}_n) =& \left(\mathrm{Var}[\bar{X}_n]\right)^{1/2} \\ =& \frac{1}{n}\left(\sum_{j=1}^{q} \mathrm{Var}[X_j'\mathbf{1}_j]\right)^{1/2} ,\end{aligned}$$

where the last equation uses that X_j are independent across clusters.

Example 18.1 In this example we assume no dependence within the cluster and we obtain a $\sqrt{n}$ convergence rate. Let us start by assuming that $n_j = n^a$ and $q = n^{1-a}$, where $a \in (0,1)$. Let $X_{i,j} \in \mathbf{R}$ be a random variable such that $\mathrm{Var}[X_{i,j}] = 1$ and $\mathrm{Cov}[X_{i,j}, X_{i',j}] = 0$. These conditions imply that

$$\mathrm{sd}(\bar{X}_n) = n^{-1/2} . \tag{18.9}$$

■

Example 18.2 In this example we assume full dependence within the cluster and we obtain a $\sqrt{q}$ convergence rate. Let us start by assuming that $n_j = n^a$ and $q = n^{1-a}$, where $a \in (0,1)$. Let $X_{i,j} \in \mathbf{R}$ be a random variable such that $\mathrm{Var}[X_{i,j}] = 1$ and $\mathrm{Cov}[X_{i,j}, X_{i',j}] = 1$. These conditions imply that

$$\mathrm{sd}(\bar{X}_n) = q^{-1/2} . \tag{18.10}$$

■

Example 18.3 In this example we assume some dependence within the cluster and we obtain a convergence rate between $q^{-1/2}$ and $n^{-1/2}$. Let us start by assuming that $n_j = n^a$ and $q = n^{1-a}$, where $a \in (0,1)$. Let $X_{i,j} \in \mathbf{R}$ be a random variable such that $\mathrm{Var}[X_{i,j}] = 1$ and $\mathrm{Cov}[X_{i,j}, X_{i',j}] = 1/|i - i'|$ for $i \neq i'$. These conditions imply that

$$\mathrm{Var}[X_j'\mathbf{1}_j] = \sum_{i=1}^{n_j}\sum_{i'=1}^{n_j} \mathrm{Cov}[X_{i,j}, X_{i',j}] = \sum_{i=1}^{n_j}\left(1 + \sum_{i'\neq i}^{n_j} \frac{1}{|i-i'|}\right) .$$

Now, let us rely on the following asymptotic proportional approximation

$$\sum_{i'=1}^{n_j} \frac{1}{|i-i'|} \propto \log(n_j) \propto \log(n) ,$$

where the last proportion approximation follows because $\log(n_j) = a\log(n)$. We can use this to conclude that

$$\mathrm{sd}(\bar{X}_n) \propto \sqrt{\frac{\log(n)}{n}} . \tag{18.11}$$

This implies that the convergence rate in this case is slower than $n^{-1/2}$, since

$$\sqrt{n}\,\mathrm{sd}(\bar{X}_n) \propto \sqrt{\log(n)} \to \infty \quad \text{as} \quad n \to \infty \ ,$$

and also implies that the convergence rate is faster than $q^{-1/2}$, since

$$\sqrt{q}\,\mathrm{sd}(\bar{X}_n) \propto \sqrt{n^{1-a}}\sqrt{\frac{\log(n)}{n}} = \sqrt{\frac{\log(n)}{n^a}} \to 0 \quad \text{as} \quad n \to \infty \ .$$

■

Example 18.4 In this example we assume full dependence within the cluster, many small clusters and few large cluster. We obtain a convergence rate slower than $q^{-1/2}$ and $n^{-1/2}$. Let us start by assuming that there are two type of clusters. In the first group, the number of cluster is $q_1 = n/2$ and $n_j = 1$ for $j = 1, \ldots, q_1$. In the second group, the number of cluster is $q_2 = n^{1-a}/2$ and $n_j = n^a$ for $j = q_1 + 1, \ldots, q_1 + q_2$, where $a \in (0, 1)$. The number of cluster is denoted by $q = q_1 + q_2$. Let $X_{i,j} \in \mathbf{R}$ be a random variable such that $\mathrm{Var}[X_{i,j}] = 1$ and $\mathrm{Cov}[X_{i,j}, X_{i',j}] = 1$. These conditions imply that

$$\mathrm{Var}[X_j'\mathbf{1}_j] = \sum_{i=1}^{n_j}\sum_{i'=1}^{n_j}\mathrm{Cov}[X_{i,j}, X_{i',j}] = n_j^2 \ ,$$

which we can use to compute the standard deviation of the sample mean,

$$\mathrm{sd}(\bar{X}_n) = \frac{1}{n}\left(\sum_{j=1}^{q} n_j^2\right)^{1/2} = \frac{1}{n}\left(q_1 + q_2 n^{2a}\right)^{1/2} \ ,$$

where the last equality follows because $\mathrm{Var}[X_j'\mathbf{1}_j] = 1$ if $j = 1, \ldots, q_1$, and $\mathrm{Var}[X_j'\mathbf{1}_j] = n^{2a}$ if $j = q_1 + 1, \ldots, q_1 + q_2$. Now, we can use that $q_1 = n/2$ and $q_2 = n^{1-a}/2$ to conclude that

$$\mathrm{sd}(\bar{X}_n) = \frac{1}{n}\left(\frac{n}{2} + \frac{n^{1-a}}{2}n^{2a}\right)^{1/2} = \left(\frac{1+n^a}{2n}\right)^{1/2} \propto n^{-(1-a)/2} \ .$$

This implies that the convergence rate in this case is slower than $n^{-1/2}$, since

$$\sqrt{n}\,\mathrm{sd}(\bar{X}_n) \propto n^{1/2}n^{-1/2}n^{a/2} \to \infty \quad \text{as} \quad n \to \infty \ ,$$

and also implies that the convergence rate is slower than $q^{-1/2}$, since $q = q_1 + q_2 \propto n$ and

$$\sqrt{q}\,\mathrm{sd}(\bar{X}_n) \propto n^{1/2}n^{-1/2}n^{a/2} \to \infty \quad \text{as} \quad n \to \infty \ .$$

This means that $\mathrm{sd}(\bar{X}_n)$ goes to zero at a slower rate than $n^{-1/2}$ and $q^{-1/2}$.
■

The final example illustrates the importance of considering heterogeneous cluster sizes. The reason why the convergence rate is slower than both $n^{-1/2}$ and $q^{-1/2}$ is because the number of clusters is determined by the large number of small clusters (q_1), but the convergence rate is determined by the (relatively) small number of large clusters ($q_2 n^{2a}$). Taken all together, these examples provide insights about the convergence rate of $\text{sd}(\bar{X}_n)$ and illustrate how this rate can be equal to $\sqrt{n}$ (square root of the sample size), equal to $\sqrt{q}$ (the square root of the number of clusters), or in-between these two rates. Most notably, under enough heterogeneity, the rate can actually be even slower than both. When $\text{sd}(\bar{X}_n)$ is a vector, it is possible that each of its elements converges at different rates.

18.8 Concluding Remarks

The notes today are taken from multiple papers. The results on the LLNs and CLTs for clustered data have been recently developed by Hansen and Lee [2019]. The wild bootstrap for clustered data was popularized by Cameron et al. [2008], but the formal properties of the method were unknown until the results in Canay et al. [2021]. Finally, the potential of randomization tests for inference with few clusters was originally studied by Canay et al. [2017].

18.9 Problems

Problem 18.1 *Show equation* (18.9) *by first showing that* $Var[X_j' \mathbf{1}_j] = n_j$.

Problem 18.2 *Show equation* (18.10) *by first showing that* $Var[X_j' \mathbf{1}_j] = n^{2a}$.

Problem 18.3 *Show equation* (18.11) *by first showing that*

$$Var[X_j' \mathbf{1}_j] \propto n_j \log(n) \ .$$

Problem 18.4 *For each statement, say whether it is true or false and justify your answer in one or two sentences, referring to the relevant assumption or result from the chapter.*

(a) The wild bootstrap test of Canay et al. [2021] is asymptotically valid for any fixed number of clusters $q \geq 2$*, with no restriction on heterogeneity across clusters.*

(b) Including cluster-level fixed effects in $W_{i,j}$ *can help the homogeneity requirement (Assumption H) hold.*

(c) When $n_j = 1$ for all clusters $j = 1, \ldots, q$, the cluster covariance estimator $\widehat{\mathbb{V}}_n$ coincides with the heteroskedasticity-robust (HC0) estimator.

(d) In the framework of Abadie et al. [2023], if treatment is assigned i.i.d. at the unit level and only a negligible fraction of population clusters is sampled, then clustering the standard errors is still required.

For (a), state both the role of Assumption H and what the lower bound $\alpha - 2^{-(q-1)}$ in Theorem 18.4 tells you about small q.

Bibliography

A. Abadie, S. Athey, G. W. Imbens, and J. M. Wooldridge. When should you adjust standard errors for clustering? *The Quarterly Journal of Economics*, 138(1):1–35, 2023.

R. M. Bell and D. F. McCaffrey. Bias reduction in standard errors for linear regression with multi-stage samples. *Survey Methodology*, 28(2):169–182, 2002.

Y. Cai, I. A. Canay, D. Kim, and A. M. Shaikh. On the implementation of approximate randomization tests in linear models with a small number of clusters. *Journal of Econometric Methods*, 12(1):85–103, 2023. doi: doi: 10.1515/jem-2021-0030.

A. C. Cameron, J. B. Gelbach, and D. L. Miller. Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, 90(3):414–427, 2008.

I. A. Canay, J. P. Romano, and A. M. Shaikh. Randomization tests under an approximate symmetry assumption. *Econometrica*, 85(3):1013–1030, May 2017.

I. A. Canay, A. Santos, and A. M. Shaikh. The wild bootstrap with a "small" number of "large" clusters. *Review of Economics and Statistics*, 103(2): 346–363, 2021.

B. E. Hansen. Jackknife standard errors for clustered regression. *Review of Economic Studies*, 2026. Forthcoming.

B. E. Hansen and S. Lee. Asymptotic theory for clustered samples. *Journal of econometrics*, 210(2):268–290, 2019.

R. Ibragimov and U. K. Müller. t-statistic based correlation and heterogeneity robust inference. *Journal of Business & Economic Statistics*, 28(4):453–468, 2010.

Part V

Topics not covered in class

A

Properties of Least Squares

Today we review the main properties of least squares. All of the properties we cover today have been covered in Econ 480-2 and are included here for completeness, since we will refer to these properties multiple times throughout the course.

A.1 Properties of LS

Let (Y, X, U) be a random vector where Y and U take values in $\mathbf{R}$ and X takes values in $\mathbf{R}^{k+1}$. Assume further that the first component of X is a constant equal to one. Let $\beta \in \mathbf{R}^{k+1}$ be such that

$$Y = X'\beta + U \ .$$

Suppose that $E[XU] = 0$, $E[XX'] < \infty$, and that there is no perfect collinearity in X. Denote by P the marginal distribution of (Y, X). Let $(Y_1, X_1), \ldots, (Y_n, X_n)$ be an i.i.d. sample of random vectors with distribution P. Above we described estimation of β via OLS under these assumptions. We now discuss properties of the resulting estimator, $\hat{\beta}_n$, imposing stronger assumptions as needed.

A.1.1 Bias

Suppose in addition that $E[U|X] = 0$. Equivalently, assume that $E[Y|X] = X'\beta$. Under this stronger assumption,

$$E[\hat{\beta}_n] = \beta \ .$$

In fact,

$$E[\hat{\beta}_n | X_1, \ldots, X_n] = \beta \ .$$

To see this, note that

$$\hat{\beta}_n = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbb{Y} = \beta + (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}\mathbb{U} \ .$$

Hence,

$$E[\hat{\beta}_n | X_1, \ldots, X_n] = \beta + (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}E[\mathbb{U}|X_1, \ldots, X_n] \ .$$

Note for any $1 \leq i \leq n$ that

$$E[U_i|X_1, \ldots, X_n] = E[U_i|X_i] = 0 \ ,$$

where the first equality follows from the fact that X_j is independent of U_i for $i \neq j$. The desired conclusion thus follows.

A.1.2 Gauss-Markov Theorem

Suppose $E[U|X] = 0$ and that $\mathrm{Var}[U|X] = \sigma^2$. When $\mathrm{Var}[U|X]$ is constant (and therefore does not depend on X) we say that U is *homoskedastic*. Otherwise, we say that U is *heteroskedastic*. The *Guass-Markov Theorem* says that under these assumptions the OLS estimator is "best" in the sense that it has the "smallest" value of $\mathrm{Var}[\mathbb{A}'\mathbb{Y}|X_1, \ldots, X_n]$ among all estimators of the form

$$\mathbb{A}'\mathbb{Y}$$

for some matrix $\mathbb{A} = \mathbb{A}(X_1, \ldots, X_n)$ satisfying

$$E[\mathbb{A}'\mathbb{Y}|X_1, \ldots, X_n] = \beta \ .$$

Here, "smallest" is understood to be in terms of the partial order obtained by defining $B \geq \tilde{B}$ if and only if $B - \tilde{B}$ is positive semi-definite. This class of estimators, of course, includes the OLS estimator as a special case (by setting $\mathbb{A}' = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'$). This property is sometimes expressed as saying that OLS is the "best linear unbiased estimator (BLUE)" of β under these assumptions.

To establish this property of OLS, first note that

$$E[\mathbb{A}'\mathbb{Y}|X_1, \ldots, X_n] = \mathbb{A}'\mathbb{X}\beta + \mathbb{A}'E[\mathbb{U}|X_1, \ldots, X_n] \ ,$$

so $E[\mathbb{A}'\mathbb{Y}|X_1, \ldots, X_n] = \beta$ if and only if $\mathbb{A}'\mathbb{X} = \mathbb{I}$. Next, note that

$$\begin{aligned} \mathrm{Var}[\mathbb{A}'\mathbb{Y}|X_1, \ldots, X_n] &= \mathbb{A}'\mathrm{Var}[\mathbb{Y}|X_1, \ldots, X_n]\mathbb{A} \\ &= \mathbb{A}'\mathrm{Var}[\mathbb{U}|X_1, \ldots, X_n]\mathbb{A} \\ &= \mathbb{A}'\mathbb{A}\sigma^2 \ . \end{aligned}$$

When $\mathbb{A}' = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'$, this last expression is simply $(\mathbb{X}'\mathbb{X})^{-1}\sigma^2$. It therefore suffices to show that

$$\mathbb{A}'\mathbb{A} - (\mathbb{X}'\mathbb{X})^{-1}$$

is positive semi-definite for all matrices $\mathbb{A}$ satisfying $\mathbb{A}'\mathbb{X} = \mathbb{I}$. To this end, define

$$\mathbb{C} = \mathbb{A} - \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1} \ .$$

Then,

$$\begin{aligned} \mathbb{A}'\mathbb{A} - (\mathbb{X}'\mathbb{X})^{-1} &= (\mathbb{C} + \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1})'(\mathbb{C} + \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}) - (\mathbb{X}'\mathbb{X})^{-1} \\ &= \mathbb{C}'\mathbb{C} + \mathbb{C}'\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1} + (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbb{C} \\ &= \mathbb{C}'\mathbb{C} \ , \end{aligned}$$

where the last equality follows from the fact that

$$\mathbb{X}'\mathbb{C} = \mathbb{X}'\mathbb{A} - \mathbb{X}'\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1} = \mathbb{I} - \mathbb{I} = 0 \ .$$

The desired conclusion thus follows from the fact that $\mathbb{C}'\mathbb{C}$ is positive semi-definite by construction.

A.1.3 Consistency

In this case we do not need additional assumptions. Note that $E[XY] < \infty$ since $XY = XX'\beta + XU$, and both $E[XX']$ and $E[XU]$ exist. Under this assumption, the OLS estimator, $\hat{\beta}_n$ is consistent for β, i.e., $\hat{\beta}_n \xrightarrow{P} \beta$ as $n \to \infty$. To see this, simply note that by the WLLN

$$\begin{aligned} \frac{1}{n} \sum_{1 \leq i \leq n} X_i X_i' &\xrightarrow{P} E[XX'] \\ \frac{1}{n} \sum_{1 \leq i \leq n} X_i Y_i &\xrightarrow{P} E[XY] \end{aligned}$$

as $n \to \infty$. The desired result therefore follows from the CMT.

A.1.4 Limiting Distribution

Suppose $E[XX'] < \infty$ and that $\mathrm{Var}[XU] = E[XX'U^2] < \infty$. Then,

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} N(0, \mathbb{V})$$

as $n \to \infty$, where

$$\mathbb{V} = E[XX']^{-1} E[XX'U^2] E[XX']^{-1} \ .$$

To see this, note that

$$\sqrt{n}(\hat{\beta}_n - \beta) = \left(\frac{1}{n} \sum_{1 \leq i \leq n} X_i X_i' \right)^{-1} \left(\frac{1}{\sqrt{n}} \sum_{1 \leq i \leq n} X_i U_i \right) \ .$$

The WLLN implies that

$$\frac{1}{n} \sum_{1 \leq i \leq n} X_i X_i' \xrightarrow{P} E[XX'] \tag{A.1}$$

as $n \to \infty$. The CLT implies that

$$\frac{1}{\sqrt{n}} \sum_{1 \leq i \leq n} X_i U_i \xrightarrow{d} N(0, \mathrm{Var}[XU])$$

as $n \to \infty$. Thus, the desired result follows from the CMT.

A.2 Estimation of $\mathbb{V}$

In order to make use of the preceding estimators, we will require a consistent estimator of

$$\mathbb{V} = E[XX']^{-1}E[XX'U^2]E[XX']^{-1} \ .$$

Note that $\mathbb{V}$ has the so-called sandwich form. As with most sandwich estimators, the interesting object is the "meat" and not the "bread". Indeed, the bread can be consistently estimated by (A.1).

Focusing our attention to the meat, we first consider the case where $E[U|X] = 0$ and $\mathrm{Var}[U|X] = \sigma^2$ (i.e., under homoskedasticity). Under these conditions,

$$\mathrm{Var}[XU] = E[XX'U^2] = E[XX']\sigma^2 \ .$$

Hence,

$$\mathbb{V} = E[XX']^{-1}\sigma^2 \ .$$

A natural choice of estimator is therefore

$$\hat{\mathbb{V}}_n = \left(\frac{1}{n}\sum_{1\leq i\leq n} X_iX_i'\right)^{-1} \hat{\sigma}_n^2 \ ,$$

where $\hat{\sigma}_n^2$ is a consistent estimator of σ^2. The main difficulty in showing that this estimator is a consistent estimator of $\mathbb{V}$ lies in choosing a consistent estimator of σ^2. A natural choice of such an estimator is

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{1\leq i\leq n} \hat{U}_i^2 \ .$$

Note that

$$\hat{U}_i = Y_i - X_i'\hat{\beta}_n = U_i - X_i'(\hat{\beta}_n - \beta) \ ,$$

so

$$\hat{U}_i^2 = (U_i - X_i'(\hat{\beta}_n - \beta))^2 = U_i^2 - 2U_iX_i'(\hat{\beta}_n - \beta) + (X_i'(\hat{\beta}_n - \beta))^2 \ .$$

The WLLN implies that

$$\frac{1}{n}\sum_{1\leq i\leq n} U_i^2 \xrightarrow{P} \sigma^2$$

as $n \to \infty$. Next, note that the WLLN and CMT imply further that

$$\frac{1}{n}\sum_{1\leq i\leq n} U_iX_i'(\hat{\beta}_n - \beta) = (\hat{\beta}_n - \beta)'\frac{1}{n}\sum_{1\leq i\leq n} X_iU_i = o_P(1) \ .$$

Finally, note that

$$\begin{aligned}\left|\frac{1}{n}\sum_{1\leq i\leq n}(X_i'(\hat{\beta}_n-\beta))^2\right| &\leq \frac{1}{n}\sum_{1\leq i\leq n}|(X_i'(\hat{\beta}_n-\beta))^2| \\ &\leq \frac{1}{n}\sum_{1\leq i\leq n}|X_i|^2|\hat{\beta}_n-\beta|^2\ ,\end{aligned}$$

which tends in probability to zero because of the WLLN, CMT and the fact that $E[|X|^2]<\infty$ (which follows from the fact that $E[XX']<\infty$). The desired conclusion thus follows.

When we do not assume $\text{Var}[U|X]=\sigma^2$, a natural choice of estimator is

$$\hat{\mathbb{V}}_n=\left(\frac{1}{n}\sum_{1\leq i\leq n}X_iX_i'\right)^{-1}\left(\frac{1}{n}\sum_{1\leq i\leq n}X_iX_i'\hat{U}_i^2\right)\left(\frac{1}{n}\sum_{1\leq i\leq n}X_iX_i'\right)^{-1}\ . \quad \text{(A.2)}$$

This estimator is consistent without additional assumptions, i.e.,

$$\hat{\mathbb{V}}_n\xrightarrow{P}\mathbb{V}\text{ as }n\to\infty\ ,$$

regardless of the functional form of $\text{Var}[U|X]$. You will find a proof in Section A.4 below. The estimator in (A.2) is called the Heteroskedasticity Consistent (HC) estimator of $\mathbb{V}$. The standard errors used to construct t-statistics are the square roots of the diagonal elements of $\hat{\mathbb{V}}_n$. It is important to note that, by default, `Stata` reports homoskedastic-only standard errors.

A.3 Improving finite sample performance: HC2 & HC3

`Stata` does not compute $\hat{\mathbb{V}}_n$ in the default "robust" option, but rather a version of this estimator that includes a finite sample adjustment to "inflate" the estimated residuals (known to be too small in finite samples). This version of the HC estimator is commonly known as $HC1$ and given by

$$\hat{\mathbb{V}}_{\text{hc1,n}}=\left(\frac{1}{n}\sum_{1\leq i\leq n}X_iX_i'\right)^{-1}\left(\frac{1}{n}\sum_{1\leq i\leq n}X_iX_i'\hat{U}_i^{*2}\right)\left(\frac{1}{n}\sum_{1\leq i\leq n}X_iX_i'\right)^{-1}\ ,$$

where $\hat{U}_i^{*2}=\frac{n}{n-k-1}\hat{U}_i^2$. It is immediate to see that this estimator is also consistent for $\mathbb{V}$. With the obvious modification for the components β_j, $1\leq j\leq k$, and using HC1 standard errors, these are the "robust" confidence intervals reported by `Stata`. Other versions, including the one discussed in the next section are also available as an option.

The consistency of the standard errors does not necessarily translate into accurate finite sample inference on β in general, something that lead to a number of finite sample adjustments that are sometimes used in practice. The simplest one is the HC1 correction, although better alternatives are available.

An alternative to $\hat{\mathbb{V}}_n$ and $\hat{\mathbb{V}}_{\text{hc1,n}}$ is what MacKinnon and White [1985] call the HC2 variance estimator, here denoted by $\hat{\mathbb{V}}_{\text{hc2,n}}$. In order to define this estimator, we need additional notation. Let

$$\mathbb{P} = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'$$

be the $n \times n$ projection matrix, with i-th column denoted by

$$P_i = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}X_i$$

and (i,i)-th element denoted by

$$P_{ii} = X_i'(\mathbb{X}'\mathbb{X})^{-1}X_i \ .$$

Let Ω be the $n \times n$ diagonal matrix with i-th diagonal element equal to $\sigma^2(X_i) = \text{Var}[U_i|X_i]$, and let $e_{n,i}$ be the n-vector with i-th element equal to one and all other elements equal to zero. Let $\mathbb{I}$ be the $n \times n$ identity matrix and $\mathbb{M} = \mathbb{I} - \mathbb{P}$ be the residual maker matrix. The residuals $\hat{U}_i = Y_i - X_i'\hat{\beta}_n$ can be written as

$$\hat{U}_i = e_{n,i}'\mathbb{M}\mathbb{U} \ , \text{ or, in vector form, } \hat{\mathbb{U}} = \mathbb{M}\mathbb{U} \ . \tag{A.3}$$

The (conditional) expected value of the square of the residual is

$$\begin{aligned} E[\hat{U}_i^2|X_1,\ldots,X_n] &= E[(e_{n,i}'\mathbb{M}\mathbb{U})^2|X_1,\ldots,X_n] \\ &= (e_{n,i} - P_i)'\Omega(e_{n,i} - P_i) \ . \end{aligned}$$

If we further assume homoskedasticity (i.e., $\text{Var}[U|X] = \sigma^2$), the last expression reduces to

$$E[\hat{U}_i^2|X_1,\ldots,X_n] = \sigma^2(1 - P_{ii}) \ ,$$

by exploiting that $\mathbb{P}$ is an idempotent matrix. In other words, even when the error term U is homoskedastic, the LS residual $\hat{U}$ is heteroskedastic (due to the presence of P_{ii}). Moreover, since it can be shown that $\frac{1}{n} \leq P_{ii} \leq 1$, it follows that $\text{Var}[\hat{U}_i]$ underestimates σ^2 under homoskedasticity. This discussion makes it natural to consider

$$\tilde{U}_i^2 \equiv \frac{\hat{U}_i^2}{1 - P_{ii}} \ , \tag{A.4}$$

as the squared residual to use in variance estimation as $\tilde{U}_i^2$ is unbiased for $E[U_i^2|X_1,\ldots,X_n]$ under homoskedasticity. This is the motivation for the variance estimator MacKinnon and White (1985) introduce as HC2,

$$\hat{\mathbb{V}}_{\text{hc2,n}} = \left(\frac{1}{n}\sum_{i=1}^n X_iX_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n X_iX_i'\tilde{U}_i^2\right)\left(\frac{1}{n}\sum_{i=1}^n X_iX_i'\right)^{-1} \ , \tag{A.5}$$

where $\tilde{U}_i^2$ is as in (A.4). Under heteroskedasticity this estimator is unbiased only in some simple examples (we will cover one of these next class), but it is biased in general. However, it is expected to have lower bias relative to HC/HC1 - a statement supported by simulations.

There are other finite sample adjustments that give place to HC3, HC4, and even HC5. For example, HC3 is equivalent to HC2 with

$$\tilde{U}_i^{*2} \equiv \frac{\hat{U}_i^2}{(1-P_{ii})^2} \ , \tag{A.6}$$

replacing $\tilde{U}_i^2$, and its justification is related to the Jackknife estimator of the variance of $\hat{\beta}_n$. It is worth noting that HC2 and HC3 are available as an option in `Stata`.

A.4 Consistency of HC standard errors

We now prove that $\hat{\mathbb{V}}_n \xrightarrow{P} \mathbb{V}$. The main difficulty lies in showing that

$$\frac{1}{n}\sum_{1\leq i\leq n} X_i X_i' \hat{U}_i^2 \xrightarrow{P} \text{Var}[XU]$$

as $n \to \infty$.

Note that

$$\frac{1}{n}\sum_{1\leq i\leq n} X_i X_i' \hat{U}_i^2 = \frac{1}{n}\sum_{1\leq i\leq n} X_i X_i' U_i^2 + \frac{1}{n}\sum_{1\leq i\leq n} X_i X_i' (\hat{U}_i^2 - U_i^2) \ .$$

Under the assumption that $\text{Var}[XU] < \infty$, the first term on the righthand side of the preceding display converges in probability to $\text{Var}[XU]$. It therefore suffices to show that the second term on the righthand side of the preceding display converges in probability to zero. We argue this separately for each of the $(k+1)^2$ terms. To this end, note for any $0 \leq j \leq k$ and $0 \leq j' \leq k$ that

$$\begin{aligned} \left| \frac{1}{n}\sum_{1\leq i\leq n} X_{i,j} X_{i,j'} (\hat{U}_i^2 - U_i^2) \right| &\leq \frac{1}{n}\sum_{1\leq i\leq n} |X_{i,j} X_{i,j'}| |\hat{U}_i^2 - U_i^2| \\ &\leq \frac{1}{n}\sum_{1\leq i\leq n} |X_{i,j} X_{i,j'}| \max_{1\leq i\leq n} |\hat{U}_i^2 - U_i^2| \ . \end{aligned}$$

Because $E[XX'] < \infty$, we have that $E[|X_j X_{j'}|] < \infty$. Hence,

$$\frac{1}{n}\sum_{1\leq i\leq n} |X_{i,j} X_{i,j'}| = O_P(1) \ ,$$

so it suffices to show that

$$\max_{1\leq i\leq n} |\hat{U}_i^2 - U_i^2| = o_P(1) \ .$$

For this purpose, the following lemma will be useful:

Lemma A.1 *Let $Z_1, \ldots, Z_n$ be an i.i.d. sequence of random vectors such that $E[|Z_i|^r] < \infty$. Then $\max_{1\leq i\leq n} |Z_i| = o_P(n^{\frac{1}{r}})$, i.e.,*

$$n^{-\frac{1}{r}} \max_{1\leq i\leq n} |Z_i| \xrightarrow{P} 0 \ .$$

PROOF: Let $\epsilon > 0$ be given. Note that

$$\begin{aligned} P\{n^{-\frac{1}{r}} \max_{1\leq i\leq n} |Z_i| > \epsilon\} &= P\{\bigcup_{1\leq i\leq n} \{|Z_i|^r > \epsilon^r n\}\} \\ &\leq \sum_{1\leq i\leq n} P\{|Z_i|^r > \epsilon^r n\} \\ &\leq \frac{1}{n\epsilon^r} \sum_{1\leq i\leq n} E[|Z_i|^r I\{|Z_i|^r > \epsilon^r n\}] \\ &= \frac{1}{\epsilon^r} E[|Z_i|^r I\{|Z_i|^r > \epsilon^r n\}] \\ &\to 0 \end{aligned}$$

as $n \to \infty$, where the first equality follows by inspection, the first inequality follows from Bonferonni's inequality, the second inequality follows from Markov's inequality, the final equality follows from the i.i.d. assumption, and the convergence to zero follows from the assumption that $E[|Z_i|^r] < \infty$. ■

We now use Lemma A.1 to establish the desired convergence in probability to zero. Note that $E[|X|^2] < \infty$ (which follows from the fact that $E[XX'] < \infty$) and $E[|UX|^2] < \infty$ (which follows from the fact that $\mathrm{Var}[XU] < \infty$). Recall that $\hat{U}_i = U_i - X_i'(\hat{\beta}_n - \beta)$, so that

$$|\hat{U}_i^2 - U_i^2| \leq 2|U_i||X_i||\hat{\beta}_n - \beta| + |X_i|^2|\hat{\beta}_n - \beta|^2 \ .$$

Next, note that Lemma A.1 and the fact that $\sqrt{n}(\hat{\beta}_n - \beta) = O_P(1)$ imply that

$$\begin{aligned} |\hat{\beta}_n - \beta| \max_{1\leq i\leq n} |U_i||X_i| &= o_P(1) \\ |\hat{\beta}_n - \beta|^2 \max_{1\leq i\leq n} |X_i|^2 &= o_P(1) \ . \end{aligned}$$

The desired conclusion thus follows. If we combine this result with

$$\frac{1}{n} \sum_{1\leq i\leq n} X_i X_i' \xrightarrow{P} E[XX'] \ ,$$

it follows immediately that

$$\hat{\mathbb{V}}_n \xrightarrow{P} \mathbb{V} \ .$$

A.5 Measures of Fit

When reporting the results of estimating a linear regression via OLS, it is common to report a measure of fit known as R^2, defined as follows:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS} \ ,$$

where

$$\begin{aligned} TSS &= \sum_{1 \leq i \leq n} (Y_i - \bar{Y}_n)^2 \\ ESS &= \sum_{1 \leq i \leq n} (\hat{Y}_i - \bar{Y}_n)^2 \\ SSR &= \sum_{1 \leq i \leq n} \hat{U}_i^2 \ . \end{aligned}$$

Here, TSS is short for *total sum of squares*, ESS is short for *explained sum of squares*, and SSR is short for *sum of squared residuals*. To show that the two expressions for R^2 are the same, and that $0 \leq R^2 \leq 1$, it suffices to show that

$$SSR + ESS = TSS \ .$$

Moreover, $R^2 = 1$ if and only if $SSR = 0$, i.e., $\hat{U}_i = 0$ for all $1 \leq i \leq n$. Similarly, $R^2 = 0$ if and only if $ESS = 0$, i.e., $\hat{Y}_i = \bar{Y}_n$ for all $1 \leq i \leq n$. In this sense, R^2 is a measure of the "fit" of a regression.

Note that $\frac{1}{n} \sum_{1 \leq i \leq n} (Y_i - \bar{Y}_n)^2$ may be viewed as an estimator of $\mathrm{Var}[Y_i]$ and $\frac{1}{n} \sum_{1 \leq i \leq n} \hat{U}_i^2$ may be viewed as an estimator of $\mathrm{Var}[U_i]$. Thus, R^2 may be viewed as an estimator of the quantity

$$1 - \frac{\mathrm{Var}[U_i]}{\mathrm{Var}[Y_i]} \ .$$

Replacing these estimators with their unbiased counterparts yields "adjusted" R^2, defined as

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1} \frac{SSR}{TSS} \ .$$

Note that R^2 always increases with the inclusion of an additional regressor, whereas $\bar{R}^2$ may not. Note also that $\bar{R}^2 \leq R^2$, so $\bar{R}^2 \leq 1$, but, unlike R^2, $\bar{R}^2$ may be less than zero.

It is important to understand that a high R^2 does not justify interpreting a linear regression as a causal model, just as a low R^2 does not invalidate interpreting a linear regression as a causal model.

A.6 Concluding Remarks

These notes are based on past notes for Econ 480-3 that I have written for previous editions of this class, and are heavily influenced by Azeem Shaikh, the notes he kindly shared with me, and our conversations about teaching over the years. Other sources that contain useful related concepts include the books by Bruce Hansen, Hansen [2022], and the one by Angrist and Pischke, Angrist and Pischke [2008]. The results in Angrist [1998] are related to recent results on treatment effects with delayed outcomes, **?**, and contamination bias when A is not binary, Goldsmith-Pinkham et al. [2022].

A.7 Problems

Problem A.1 *Prove* (3.5)

Problem A.2 *Complete all of the steps required to go from* (3.4) *to* (3.8)

Problem A.3 *Consider the characterization of* β_A *in* (3.8)*. Show that*

$$\beta_A = \sum_{w \in \mathcal{W}} \omega(w)\Delta(w) \ , \tag{A.7}$$

where the weights $\omega(w)$ *are non-random and satisfy* $\omega(w) \geq 0$ *for all* $w \in \mathcal{W}$ *and* $\sum_{w \in \mathcal{W}} \omega(w) = 1$*.*

Problem A.4 *Using Angrist(1998)'s dataset, do the following regressions separately for the whites and non-whites:*

1. *regression of earnings on the veteran status only, i.e., the difference in means estimator;*
2. *regression saturated in discrete covariates as in* (3.2)*, only including the year of birth as the covariate;*
3. *regression fully-saturated in discrete covariates as in* (3.3)*, only including the year of birth as the covariate;*

Report and compare your estimates for β_A *with column (2) and (4) in Table 3.1. The variables in the dataset you may use are the following:*

- *DNWHITE: dummy variable, 0 if white, 1 otherwise.*
- *DVET: dummy variable, 1 if veteran, 0 otherwise.*

- *DOBYY: year of birth.*
- *EARNVAR: earnings.*
- *YEAR: the year corresponding to the earnings record. (To replicate Table 3.1, you may select the years 1988-1991.)*

Bibliography

J. D. Angrist. Estimating the labor market impact of voluntary military service using social security data on military applicants. *Econometrica*, 66 (2):249–288, 1998.

J. D. Angrist and J.-S. Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2008.

P. Goldsmith-Pinkham, P. Hull, and M. Kolesár. Contamination bias in linear regressions. Technical report, National Bureau of Economic Research, 2022.

B. Hansen. *Econometrics*. Princeton University Press, 2022.

J. G. MacKinnon and H. White. Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, 29(3):305–325, 1985.

B

Basic Inference

B.1 Inference

Let (Y, X, U) be a random vector where Y and U take values in $\mathbf{R}$ and X takes values in $\mathbf{R}^{k+1}$. Assume further that the first component of X is a constant equal to one. Let $\beta \in \mathbf{R}^{k+1}$ be such that

$$Y = X'\beta + U \ .$$

Suppose that $E[XU] = 0$, that there is no perfect collinearity in X, that $E[XX'] < \infty$, and $\mathrm{Var}[XU] < \infty$. Denote by P the marginal distribution of (Y, X). Let $(Y_1, X_1), \ldots, (Y_n, X_n)$ be an i.i.d. sample of random vectors with distribution P. Under these assumptions, we established the asymptotic normality of the OLS estimator $\hat{\beta}_n$,

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} N(0, \mathbb{V}) \tag{B.1}$$

with

$$\mathbb{V} = E[XX']^{-1} E[XX'U^2] E[XX']^{-1} \ . \tag{B.2}$$

We also described a consistent estimator $\hat{\mathbb{V}}_n$ of the limiting variance $\mathbb{V}$. We now use these results to develop methods for inference. We will study in particular *Wald tests* for certain hypotheses. Some other testing principles will be covered later in class. Confidence regions will be constructed using the duality between hypothesis testing and the construction of confidence regions.

Below we will assume further that $\mathrm{Var}[XU] = E[XX'U^2]$ is non-singular. This would be implied, for example, by the assumption that $P\{E[U^2|X] > 0\} = 1$. Since $E[XX']$ is non-singular under the assumption of no perfect collinearity in X, this implies that $\mathbb{V}$ is non-singular.

B.1.1 Background

Consider the following somewhat generic version of a testing problem. One observes data $W_i = (Y_i, X_i), i = 1, \ldots, n$, i.i.d. with distribution $P \in \mathbf{P} = \{P_\beta : \beta \in \mathbf{R}^{k+1}\}$ and wishes to test

$$H_0 : \beta \in \mathbf{B}_0 \quad \text{versus} \quad H_1 : \beta \in \mathbf{B}_1 \tag{B.3}$$

where $\mathbf{B}_0$ and $\mathbf{B}_1$ form a partition of $\mathbf{R}^{k+1}$. In our context, β will be the coefficient in a linear regression but in general it could be any other parameter.

A test is simply a function $\phi_n = \phi_n(W_1, \ldots, W_n)$ that returns the probability of rejecting the null hypothesis after observing $W_1, \ldots, W_n$. For the time being, we will only consider non-randomized tests which means that the function ϕ_n will take only two values: it will be equal to 1 for rejection and equal to 0 for non rejection. Most often, ϕ_n is the indicator function of a certain test statistic $T_n = T_n(W_1, \ldots, W_n)$ being greater than some critical value $c_n(1-\alpha)$, this is,

$$\phi_n = I\left\{T_n > c_n(1-\alpha)\right\} \ . \tag{B.4}$$

The test is said to be (pointwise) asymptotically of level α (or consistent in levels) if,

$$\limsup_{n\to\infty} E_{P_\beta}\left[\phi_n\right] = \limsup_{n\to\infty} P_\beta\{\phi_n = 1\} \leq \alpha \ , \quad \forall \beta \in \mathbf{B}_0 \ .$$

Such tests include: Wald tests, quasi-likelihood ratio tests, and Lagrange multiplier tests.

B.1.2 Tests of A Single Linear Restriction

Consider testing

$$H_0 : r'\beta = c \text{ versus } H_1 : r'\beta \neq c \ ,$$

where r is a nonzero $(k+1)$-dimensional vector and c is a scalar, at level α. Probably the most important case in this class happens when r selects the sth component of β, in which case we get

$$H_0 : \beta_s = c \text{ versus } H_1 : \beta_s \neq c \ .$$

The CMT implies that

$$\sqrt{n}(r'\hat{\beta}_n - r'\beta) \xrightarrow{d} N(0, r'\mathbb{V}r)$$

as $n \to \infty$. Since $\mathbb{V}$ is non-singular, $r'\mathbb{V}r > 0$. The CMT implies that $r'\hat{\mathbb{V}}_n r \xrightarrow{P} r'\mathbb{V}r$ as $n \to \infty$. A natural choice of test statistic for this problem is the absolute value of the t-statistic,

$$t_{\text{stat}} = \frac{\sqrt{n}(r'\hat{\beta}_n - c)}{\sqrt{r'\hat{\mathbb{V}}_n r}} \ ,$$

so that $T_n = |t_{\text{stat}}|$. Note that when r selects the sth component of β, we get $r'\hat{\mathbb{V}}_n r = \hat{\mathbb{V}}_{n,[s,s]}$, i.e., the sth diagonal element of $\hat{\mathbb{V}}_n$.

Such test statistic has the property that large values of T_n provide evidence against the null hypothesis H_0, and so using rejection rules of the form "reject H_0 if T_n is greater than a certain threshold" makes sense. This threshold value is usually called *critical value*.

A suitable choice of critical value for this test statistic is $z_{1-\frac{\alpha}{2}}$, which exploits the fact that, under the null hypothesis,

$$t_{\text{stat}} = \frac{\sqrt{n}(r'\hat{\beta}_n - c)}{\sqrt{r'\hat{\mathbb{V}}_n r}} \xrightarrow{d} N(0,1) \ . \tag{B.5}$$

To see that this test is consistent in level, note that whenever $r'\beta = c$,

$$\begin{aligned} P\{\phi_n = 1\} &= P\{|T_n| > z_{1-\frac{\alpha}{2}}\} \\ &= P\{t_{\text{stat}} > z_{1-\frac{\alpha}{2}}\} + P\{t_{\text{stat}} < -z_{1-\frac{\alpha}{2}}\} \\ &\to 1 - \Phi(z_{1-\frac{\alpha}{2}}) + \Phi(-z_{1-\frac{\alpha}{2}}) \\ &= 1 - \Phi(z_{1-\frac{\alpha}{2}}) + \Phi(z_{\frac{\alpha}{2}}) \\ &= 1 - (1 - \frac{\alpha}{2}) + \frac{\alpha}{2} \\ &= \alpha \ . \end{aligned}$$

This construction may be modified in a straightforward fashion for testing "one-sided" hypotheses, i.e.,

$$H_0 : r'\beta \leq c \text{ versus } H_1 : r'\beta > c \ .$$

In addition, note that by using the duality between hypothesis testing and the construction of confidence regions, we may construct a confidence region of level α for each component β_s of β as

$$\begin{aligned} C_n &= \left\{ c \in \mathbf{R} : \left| \frac{\sqrt{n}(\hat{\beta}_{n,s} - c)}{\sqrt{\hat{\mathbb{V}}_{n,[s,s]}}} \right| \leq z_{1-\frac{\alpha}{2}} \right\} \\ &= \left\{ \hat{\beta}_{n,s} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{\mathbb{V}}_{n,[s,s]}}{n}}, \hat{\beta}_{n,s} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{\mathbb{V}}_{n,[s,s]}}{n}} \right\} \ . \end{aligned}$$

This confidence region satisfies

$$P\{\beta_s \in C_n\} \to 1 - \alpha$$

as $n \to \infty$. It is straightforward to modify this construction to construct a confidence region of level α for $r'\beta$.

B.1.3 Tests of Multiple Linear Restrictions

Consider testing

$$H_0 : R\beta = c \text{ versus } H_1 : R\beta \neq c \ ,$$

where R is a $p \times (k+1)$-dimensional matrix and c is a p-dimensional vector, at level α. In order to rule out redundant equations, we require that the rows of R are linearly independent. The CMT implies that

$$\sqrt{n}(R\hat{\beta}_n - R\beta) \xrightarrow{d} N(0, R\mathbb{V}R')$$

as $n \to \infty$. Note that because $\mathbb{V}$ is assumed to be non-singular, $R\mathbb{V}R'$ is also non-singular. To see this, consider $a'R\mathbb{V}R'a$ for a non-zero vector $a \in \mathbf{R}^p$. Next, note that $a'R \neq 0$ because the rows of R are assumed to be linearly independent. Hence, $a'R\mathbb{V}R'a > 0$ because $\mathbb{V}$ is assumed to be non-singular. Hence, from our earlier results, we see that

$$n(R\hat{\beta}_n - R\beta)'(R\hat{\mathbb{V}}_n R')^{-1}(R\hat{\beta}_n - R\beta) \xrightarrow{d} \chi^2_p$$

as $n \to \infty$. Thus, a natural choice of test statistic in this case is therefore

$$T_n = n(R\hat{\beta}_n - c)'(R\hat{\mathbb{V}}_n R')^{-1}(R\hat{\beta}_n - c)$$

and a suitable choice of critical value is $c_{p,1-\alpha}$, the $1-\alpha$ quantile of χ^2_p. The resulting test is consistent in level.

Note that by using the duality between hypothesis testing and the construction of confidence regions, we may construct a confidence region of level α for β as

$$C_n = \{c \in \mathbf{R}^{k+1} : n(\hat{\beta}_n - c)'\hat{\mathbb{V}}_n^{-1}(\hat{\beta}_n - c) \leq c_{k+1,1-\alpha}\} \ .$$

This confidence region satisfies

$$P\{\beta \in C_n\} \to 1 - \alpha$$

as $n \to \infty$. It is straightforward to modify this construction to construct a confidence region of level α for $R\beta$.

B.1.4 Tests of Nonlinear Restrictions

Consider testing

$$H_0 : f(\beta) = 0 \text{ versus } H_1 : f(\beta) \neq 0 \ ,$$

where $f : \mathbf{R}^{k+1} \to \mathbf{R}^p$, at level α. Assume that f is continuously differentiable at β and denote by $D_\beta f(\beta)$ the $p \times (k+1)$-dimensional matrix of partial derivatives of f evaluated at β. Assume that the rows of $D_\beta f(\beta)$ are linearly independent. The Delta Method implies that

$$\sqrt{n}(f(\hat{\beta}_n) - f(\beta)) \xrightarrow{d} N(0, D_\beta f(\beta)\mathbb{V}D_\beta f(\beta)')$$

as $n \to \infty$. The CMT implies that

$$D_\beta f(\hat{\beta}_n)\hat{\mathbb{V}}_n D_\beta f(\hat{\beta}_n)' \xrightarrow{P} D_\beta f(\beta)\mathbb{V}D_\beta f(\beta)'$$

as $n \to \infty$. It is now straightforward to modify the construction of the test in the preceding section appropriately to develop a test for the present purpose. It is also straightforward to modify the construction of the confidence region in the preceding section to construct a confidence region of level α for $f(\beta)$.

Bibliography

B. Hansen. *Econometrics*. Princeton University Press, 2022.

C

Additional Topics in RDD

C.1 RD Plots

An appealing feature of the RD design is that it can be illustrated graphically. This graphical representation adds transparency to the analysis by displaying the observations used for estimation and inference. RD plots also allow researchers to readily summarize the main empirical findings as well as other important features of the work conducted.

We now discuss the most transparent and effective methods to graphically illustrate the RD design with the Meyersson example. At first glance, it seems that one should be able to illustrate the relationship between the outcome and the score by simply constructing a scatter plot, clearly identifying the points above and below the cutoff. However, this strategy is rarely useful, as it is often hard to see "jumps" or discontinuities in the outcome-score relationship by simply looking at the raw data. We illustrate this point in Figure C.1 with the Meyersson data.

A more useful approach is to aggregate or "smooth" the data before plotting. The typical RD plot presents two summaries: (i) a global polynomial fit, represented by a solid line, and (ii) local sample means, represented by dots. The global polynomial fit is simply a smooth approximation to the unknown regression functions based on a fourth- or fifth-order polynomial regression of the outcome on the score, fitted separately above and below the cutoff, and using the original raw data. In contrast, the local sample means are created by first choosing disjoint intervals or "bins" of the score, calculating the mean of the outcome for the observations falling within each bin, and then plotting the average outcome in each bin against the mid point of the bin. Local sample means can be interpreted as a non-smooth approximation to the unknown regression functions. The combination of these two ingredients in the same plot allows researchers to visualize the global shape of the regression functions for treated and control observations, while at the same time retaining enough information about the local behavior of the data to observe the RD treatment effect and the variability of the data around the global fit.

For example, in the Meyersson application, if we use 20 bins of equal length on each side of the cutoff, we partition the support of the Islamic margin of victory into 40 disjoint intervals of length 5. In Figure C.2, we plot the binned

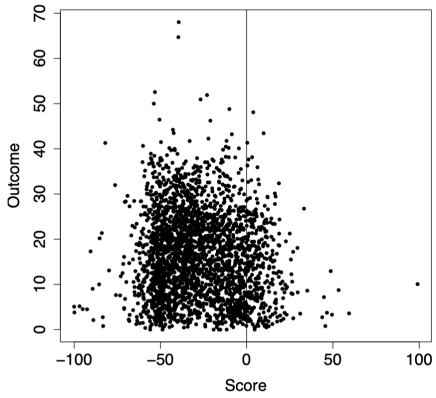


FIGURE C.1: Scatter Plot (Meyersson Data)

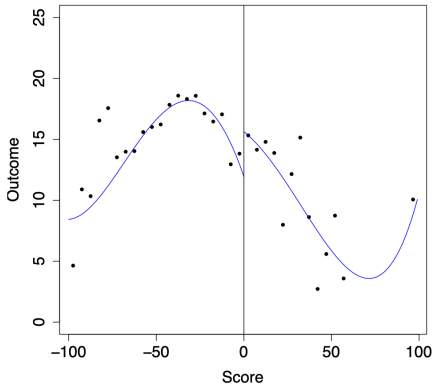


FIGURE C.2: RD Plot for Meyersson Data Using 40 Bins of Equal Length

outcome means against the score, adding a fourth-order global polynomial fit estimated separately for treated and control observations. The binned means let us see the local behavior of the average response variable around the global fit. The plot also reveals a positive jump at the cutoff: the average educational attainment of women seems to be higher in municipalities where the Islamic party barely won than where the Islamic party barely lost. Binning the data may reveal striking patterns that can remain hidden in a simple scatter plot. We now discuss how to choose the type and number of bins in a data-driven and transparent way.

C.1.1 Choosing the Location of Bins

There are two different types of bins that can be used in RD plots: bins that have equal length, or bins that contain (roughly) the same number of observations but whose length may differ. We refer to these two types as evenly-spaced (ES) and quantile-spaced (QS) bins, respectively.

In order to define the bins more precisely, we assume that the running variable takes values inside the interval $[x_l, x_u]$. Let $X_{-,q}$ and $X_{+,q}$ denote the q-th quantiles of the control and treatment subsamples, respectively, for $q \in (0, 1)$. Suppose we decide to have 10 bins for both the control and treated subsamples. Then the two types of bins are defined in the following.

- **Evenly-spaced (ES) bins:**

$$[x_l, x_l + \frac{(c - x_l)}{10}), \quad [x_l + \frac{(c - x_l)}{10}, x_l + \frac{2\,(c - x_l)}{10}), \quad ..., \quad [x_l + \frac{9\,(c - x_l)}{10}, c)$$

$$[c, c + \frac{(x_u - c)}{10}), \quad [c + \frac{(x_u - c)}{10}, c + \frac{2\,(x_u - c)}{10}), \quad ..., \quad [c + \frac{9\,(x_u - c)}{10}, x_u]$$

- **Quantile-spaced (QS) bins:**

$$[x_l, X_{-,0.1}), \quad [X_{-,0.1}, X_{-,0.2}), \quad ..., \quad [X_{-,0.9}, c)$$

$$[c, X_{+,0.1}), \quad [X_{+,0.1}, X_{+,0.2}), \quad ..., \quad [X_{+,0.9}, x_u]$$

The most important difference between ES and QS bins is the underlying variability of the local mean estimate in every bin. Although ES bins have equal length, if the observations are not uniformly distributed, each bin may contain a different number of observations. In an RD plot with ES bins, each of the local means represented by a dot may be computed using a different number of observations and thus may be more or less precisely calculated than the other local means in the plot, affecting comparability.

In contrast, QS bins contain approximately the same number of observations by construction. Moreover, a quantile-spaced RD plot has the advantage of providing a quick visual representation of the density of observations over the support of the score.

C.1.2 Choosing the Number of Bins

Once the positioning of the bins has been decided by choosing either QS or ES bins, it remains to choose the total number of bins on either side of the cutoff – denoted by the quantities J_- and J_+.

The first method we discuss selects the values of J_- and J_+ that minimize an asymptotic approximation to the integrated mean-squared error (IMSE) of the local means estimator, that is, the sum of the expansions of the (integrated) variance and squared bias. By construction, it will result in binned

sample means that "trace out" the underlying regression function; this is useful to assess the overall shape of the regression function. However, the IMSE-optimal method often results in a very smooth plot where the local means nearly overlap with the global polynomial fit, and may not be appropriate to capture the local variability of the data near the cutoff.

The IMSE-optimal values of J_- and J_+ are, respectively,

$$J_-^{\text{IMSE}} = \left\lceil \mathscr{C}_-^{\text{IMSE}} n^{1/3} \right\rceil \quad \text{and} \quad J_+^{\text{IMSE}} = \left\lceil \mathscr{C}_+^{\text{IMSE}} n^{1/3} \right\rceil,$$

where n is the total number of observations, and the exact form of the constants $\mathscr{C}_-^{\text{IMSE}}$ and $\mathscr{C}_+^{\text{IMSE}}$ depends on whether ES or QS bins are used and some features of the underlying data generating process. In practice, the unknown constants $\mathscr{C}_-^{\text{IMSE}}$ and $\mathscr{C}_+^{\text{IMSE}}$ are estimated using preliminary procedures.

The second method selects the number of bins so that the binned means have an asymptotic (integrated) variability that is approximately equal to the variability of the raw data. We refer to this choice of total number of bins as a mimicking variance (MV) choice.

The mimicking-variance values J_- and J_+ are,

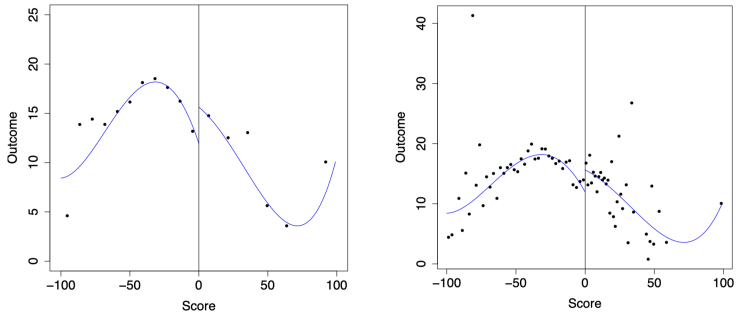
$$J_-^{\text{MV}} = \left\lceil \mathscr{C}_-^{\text{MV}} \frac{n}{\log(n)^2} \right\rceil, \quad \text{and} \quad J_+^{\text{MV}} = \left\lceil \mathscr{C}_+^{\text{MV}} \frac{n}{\log(n)^2} \right\rceil,$$

where again the constants $\mathscr{C}_-^{\text{MV}}$ and $\mathscr{C}_+^{\text{MV}}$ depend on the type of bins used and some features of the underlying data generating process. They are also estimated using preliminary procedures.

In general, the MV method leads to a larger number of bins than the IMSE method, resulting in an RD plot with more dots representing local means and thus giving a better sense of the variability of the data, as illustrated by Figure C.3. Which method of implementation is most appropriate depends on the researcher's particular goal, for example, illustrating/testing for the overall functional form versus showing the variability of the data. We recommend to start with MV bins to better illustrate the variability of the outcome as a function of the score, ideally comparing ES to QS bins to highlight the distributional features of the score. Then, if needed, the researcher can select the number of bins to be IMSE-optimal in order to explore the global features of the regression function.

C.2 Validation and Falsification of the RD Design

The identification of RDD hinges on the continuity of the regression functions $E\left[Y_i(d) \mid X_i = x\right]$ at $x = c$, yet this assumption can be violated in some situations. For example, a scholarship may be assigned based on whether students receive an exam grade above a cutoff, but the cutoff is known to the students'



(a) IMSE (b) MV

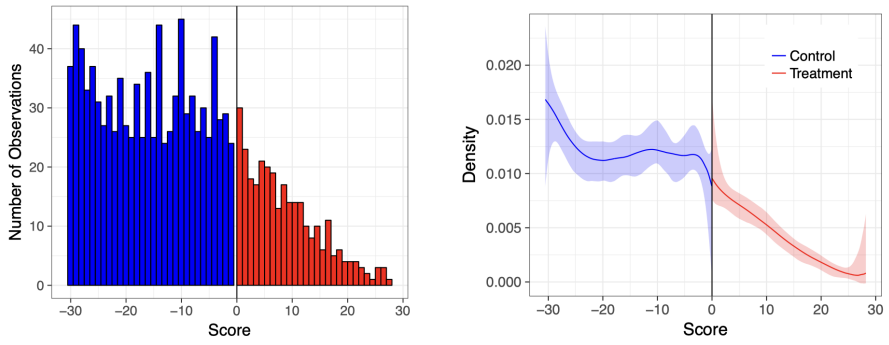
FIGURE C.3: RD Plot with Evenly-Spaced Bins (Meyersson Data)

parents and there are mechanisms to appeal the grade. If the parents who are successful in appealing the grade when their child is barely below the cutoff are systematically different from those who choose not to appeal in ways that affect the outcome of interest, then the assumption that the average potential outcomes are continuous at the cutoff is invalid. For instance, if the outcome of interest is performance on a future exam and parent involvement is positively correlated with students' future academic achievement, the average potential outcomes of students at or just above the cutoff will be higher than the average potential outcomes of students just below the cutoff.

Naturally, the continuity assumption is about unobservable features and inherently untestable. Nonetheless, there are empirical methods that can provide useful evidence about the plausibility of this assumption. These validation methods are based on various empirical implications of the unobservable RD assumption that can be expected to hold in most cases.

C.2.1 Density of Running Variable

If units were able to systematically increase the value of their original score to be assigned to the treatment instead of the control group, the density of the score right below the cutoff would be lower than just above it. To detect this type of violation of the RD assumption, McCrary [2008] suggests testing the null hypothesis that the density of the score is continuous at the cutoff. Figure C.4 provides a graphical representation of the continuity in density test approach using the Meyersson data, exhibiting both a histogram of the data and the density estimate with shaded 95% confidence intervals. The density estimates for treated and control groups at the cutoff are very near each other, and the confidence intervals (shaded areas) overlap. This means that we fail to reject the continuity of the density of the score at the cutoff, supporting the validity of the RD design.



(a) Histogram (b) Estimated Density

FIGURE C.4: Histogram and Estimated Density of the Score (Meyersson Data)

C.2.2 Predetermined Covariates and Placebo Outcomes

If units can not manipulate the score value they receive, there should be no systematic differences between units with similar values of the score. Thus, units just above and just below the cutoff should be similar in all variables that could not have been affected by the treatment. These variables can be divided into two groups: variables that are determined before the treatment is assigned – which we call predetermined covariates; and variables that are determined after the treatment is assigned but, according to substantive knowledge about the treatment's causal mechanism, could not possibly have been affected by the treatment – which we call placebo outcomes.

Note that predetermined covariates can be unambiguously defined, but placebo outcomes are always specific to each application. For example, if the treatment is access to clean water and the outcome of interest is child mortality, a treatment effect is expected on mortality due to water-borne illnesses but not on mortality due to other causes such as car accidents. Thus, mortality from road accidents would be a reasonable placebo outcome.

In practice, quite often researchers carry out this type of falsification analysis by using the following heuristic argument: all predetermined covariates and placebo outcomes should be analyzed in the same way as the outcome of interest. This means that for each predetermined covariate or placebo outcome, researchers use local polynomial techniques to estimate the "treatment effect" and employ valid inference procedures discussed previously. Since the predetermined covariate or placebo outcome could not have been affected by the treatment, the null hypothesis of no treatment effect should not be rejected if the RD design is valid. For example, Table C.1 contains the local polynomial estimation and inference results for several predetermined covariates in the Meyersson dataset. All 95% robust confidence intervals contain zero, with p-values ranging from 0.333 to 0.999. In other words, there is no

TABLE C.1: Falsification Test for Predetermined Covariates (Meyersson Data)

Variable	MSE-Optimal Bandwidth	RD Estimator	Robust Inference p-value	Robust Inference Conf. Int.	Eff. Number Observations
Percentage of men aged 15-20 with high school education	12.055	1.561	0.358	[-1.757, 4.862]	590
Islamic Mayor in 1989	11.782	0.053	0.333	[-0.077, 0.228]	418
Islamic percentage of votes in 1994	13.940	0.603	0.711	[-2.794, 4.095]	668
Number of parties receiving votes 1994	12.166	-0.168	0.668	[-1.357, 0.869]	596
Log population in 1994	13.319	0.012	0.999	[-0.644, 0.645]	633
District center	13.033	-0.067	0.462	[-0.285, 0.130]	624
Province center	11.556	0.029	0.609	[-0.064, 0.109]	574
Sub-metro center	10.360	-0.016	0.572	[-0.114, 0.063]	513
Metro center	13.621	0.008	0.723	[-0.047, 0.068]	642

TABLE C.2: RD Analysis for True and Placebo Cutoffs (Meyersson Data)

Alternative Cutoff	MSE-Optimal Bandwidth	RD Estimator	Robust Inference p-value	Robust Inference Conf. Int.	N. of Obs. Left	N. of Obs. Right
-3	3.934	1.688	0.421	$[-3.509, 8.397]$	135	74
-2	4.642	-2.300	0.991	$[-9.414, 9.518]$	152	47
-1	4.510	-3.003	0.992	$[-11.295, 11.409]$	139	24
0	17.239	3.020	0.076	$[-0.309, 6.276]$	529	266
1	2.362	-1.131	0.787	$[-9.967, 13.147]$	30	49
2	2.697	-1.973	0.488	$[-15.333, 7.313]$	53	50
3	2.850	3.766	0.668	$[-8.700, 13.569]$	68	56

empirical evidence that these predetermined covariates are discontinuous at the cutoff.

C.2.3 Placebo Cutoffs

Another falsification analysis examines treatment effects at alternative or placebo cutoff values. Evidence of continuity away from the cutoff is, of course, neither necessary nor sufficient for continuity at the cutoff, but the presence of discontinuities away from the cutoff can be interpreted as potentially casting doubt on the RD design, at the very least in cases where such discontinuities can not be explained by substantive knowledge of the specific application.

This test replaces the true cutoff value by another value at which the treatment status does not really change, and performs estimation and inference using this artificial cutoff. The expectation is that no significant treatment effect will occur.

Table C.2 presents the RD estimation and inference results for the true cutoff (0) and various placebo cutoffs. We find that all p-values for placebo cutoffs are above 0.4. Therefore, we conclude that the outcome of interest does not jump discontinuously at the placebo cutoffs considered.

TABLE C.3: RD Analysis for the Donut-Hole Approach (Meyersson Data)

Donut-Hole Radius	MSE-Optimal Bandwidth	RD Estimator	Robust Inference p-value	Robust Inference Conf. Int.	Number of Observations	Excluded Obs. Left	Excluded Obs. Right
0.00	17.239	3.020	0.076	[-0.309, 6.276]	795	0	0
0.10	17.954	3.081	0.064	[-0.175, 6.298]	815	1	1
0.20	16.621	3.337	0.052	[-0.033, 6.759]	765	5	4
0.30	16.043	3.414	0.055	[-0.067, 6.965]	730	7	6
0.40	17.164	3.286	0.050	[-0.001, 6.601]	774	9	9
0.50	15.422	3.745	0.028	[0.408, 7.292]	697	13	14

C.2.4 Sensitivity to Observations near the Cutoff

Another falsification approach investigates how sensitive the results are to the response of units who are located very close to the cutoff. If systematic manipulation of score values has occurred, it is natural to assume that the units closest to the cutoff are those most likely to have engaged in manipulation. The idea behind this approach is to exclude such units and then repeat the estimation and inference analysis using the remaining sample. This idea is sometimes referred to as a "donut hole" approach.

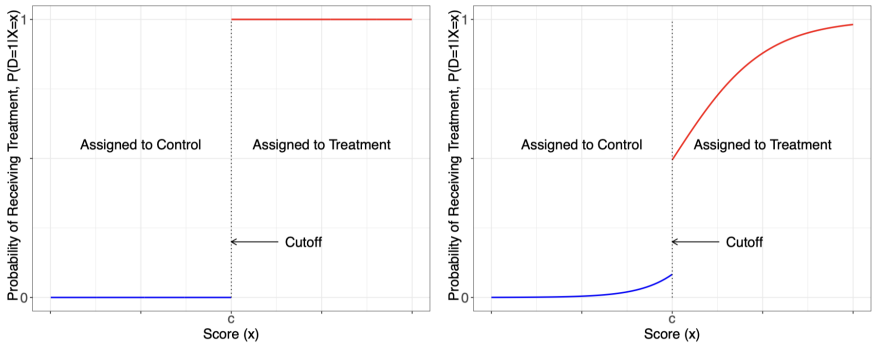
Table C.3 presents the RD estimation and inference results excluding units within different radius from the cutoff. We find that the analysis remains largely unchanged, since both the original and the new estimated effect are significant at 10% level.

C.3 The Fuzzy RD Design

It is common in practice to encounter RD designs where either some of the units with $X_i \geq c$ fail to receive the treatment or some of the units with $X_i < c$ receive the treatment anyway – or both. The phenomenon of units receiving a treatment condition different from the condition originally assigned to them is generally known as imperfect compliance or non-compliance. The RD design with imperfect compliance is usually referred to as the Fuzzy RD design.

In all RD designs, the assignment of treatment follows the rule $I\{X_i \geq c\}$, which assigns all units whose score is below the cutoff c to the control condition, and all units whose score is above c to the treatment condition. In the Sharp RD design, all units assigned to the treatment condition do in fact take the treatment, and no units assigned to the control condition take the treatment. In this case, the rule $I\{X_i \geq c\}$ indicates not only the treatment *assigned* to the units, but also the treatment *received* by the units, so we define $A_i = I\{X_i \geq c\}$.

In the Fuzzy RD, we still use the binary variable A_i to denote the treatment



(a) Sharp RD (b) Fuzzy RD

FIGURE C.5: Conditional Probability of Receiving Treatment in Sharp vs. Fuzzy RD Designs

actually received by unit i, understanding that $A_i \neq I\{X_i \geq c\}$. In order to clearly distinguish between the treatment received and the treatment assigned, we now define $Z_i := I\{X_i \geq c\}$ to be the *intention to treat*. The difference between the Sharp and Fuzzy RD designs is illustrated in Figure C.5, where we plot the conditional probability of receiving treatment given the score, $P\{A_i = 1 \mid X_i = x\}$, for different values of score.

The treatment received A_i has two potential values: $A_i(1)$ is the treatment received by i when assigned to the treatment condition (i.e, when $X_i \geq c$ or $Z_i = 1$) and $A_i(0)$ is the treatment received by i when assigned to the control condition (i.e, when $X_i < c$ or $Z_i = 0$). The observed treatment taken is

$$A_i = Z_i A_i(1) + (1 - Z_i)\, A_i(0) \ .$$

The presence of non-compliance introduces confounding between potential outcomes and compliance decisions that, in the absence of additional assumptions, prevents us from learning causal treatment effects for all units at the cutoff. Thus, it is common to shift the focus to different parameters that can still be recovered under reasonable assumptions and, despite being less general than the sharp RD treatment effect, are still of interest. These are local versions of the LATE assumptions of Imbens–Angrist; the augmented continuity assumption below is what lets us interpret the discontinuity at the cutoff as a local causal parameter. In particular, we impose the following assumptions for units near the cutoff (i.e. conditional on $X_i = x$ for x in a small neighborhood of c).

- Instrument exogeneity: $(Y_i(1), Y_i(0), A_i(1), A_i(0)) \perp\!\!\!\perp Z_i$.
- Instrument relevance: $P\{A_i(1) \neq A_i(0)\} > 0$.
- Monotonicity: $P\{A_i(1) \geq A_i(0)\} = 1$.

The instrument exogeneity assumption restricts that the treatment assignment must be independent from the potential outcomes and the potential treatments. To explain the other two assumptions, we define four groups of units according to their compliance decisions. *Compliers* are those units whose treatment received coincides with their treatment assigned; *Never-takers* are those who always refuse the treatment regardless of their assignment; *Always-takers* are those who always take the treatment regardless of their assignment; and *Defiers* are those who receive the opposite treatment to the one they are assigned. The monotonicity assumption excludes defiers. And the instrument relevance assumption requires that there must exist compliers.

In addition to the above assumptions common in the LATE literature, in the current RD setting we also need the following continuity assumption.

- Augmented continuity: $E\left[Y_i(1) \mid X_i = x\right]$, $E\left[Y_i(0) \mid X_i = x\right]$, $E\left[A_i(1) \mid X_i = x\right]$ and $E\left[A_i(0) \mid X_i = x\right]$ are continuous functions of x at $x = c$.

The Fuzzy RD parameter is defined as

$$\theta_{\text{frd}} = \frac{\lim_{x \downarrow c} E\left[Y_i \mid X_i = x\right] - \lim_{x \uparrow c} E\left[Y_i \mid X_i = x\right]}{\lim_{x \downarrow c} E\left[A_i \mid X_i = x\right] - \lim_{x \uparrow c} E\left[A_i \mid X_i = x\right]}.$$

Under the above assumptions, it can be shown that the Fuzzy RD parameter recovers the average treatment effect at the cutoff for compliers:

$$\theta_{\text{frd}} = E\left[Y_i(1) - Y_i(0) \mid X_i = c, A_i(1) > A_i(0)\right]. \tag{C.1}$$

Estimation for the Fuzzy RD parameter proceeds in the same manner as in the Sharp RD case, by simply using local polynomials (LP) on both sides of the cutoff to estimate the regression functions. This procedure needs to be done for both the numerator (Y) and denominator (A).

$$\widehat{\theta}_{\text{frd}} = \frac{\lim_{x \downarrow c} \widehat{E}\left[Y_i \mid X_i = x\right] - \lim_{x \uparrow c} \widehat{E}\left[Y_i \mid X_i = x\right]}{\lim_{x \downarrow c} \widehat{E}\left[A_i \mid X_i = x\right] - \lim_{x \uparrow c} \widehat{E}\left[A_i \mid X_i = x\right]},$$

where here $\widehat{E}$ denotes the LP estimators of the conditional expectations. Inference can also be based on robust local polynomial methods as discussed in the Sharp RD case.

C.4 Concluding Remarks

The material today heavily borrows from two books on RDD: Cattaneo et al. [2019] and Cattaneo et al. [2024], as well as slides shared by Matias Cattaneo. I want to particularly thank Matias for sharing his slides and codes with me. Other references related to validation of RDD designs include McCrary [2008], Canay and Kamat [2018], Bugni and Canay [2021].

C.5 Problems

Problem C.1 *Prove* (C.1)*. Hint: generalize the notation for the potential outcomes to* $Y_i(z, a)$*, where* z *denotes assignment and* a *denotes treatment received, and impose the exclusion restriction that assignment affects outcomes only through treatment received. Then define* $Y_i(1) = Y_i(1, 1)$ *and* $Y_i(0) = Y_i(0, 0)$*.*

Bibliography

F. A. Bugni and I. A. Canay. Testing continuity of a density via g-order statistics in the regression discontinuity design. *Journal of Econometrics*, 221(1):138–159, 2021.

I. A. Canay and V. Kamat. Approximate permutation tests and induced order statistics in the regression discontinuity design. *The Review of Economic Studies*, 85(3):1577–1608, 2018.

M. D. Cattaneo, N. Idrobo, and R. Titiunik. *A practical introduction to regression discontinuity designs: Foundations*. Cambridge University Press, 2019.

M. D. Cattaneo, N. Idrobo, and R. Titiunik. *A practical introduction to regression discontinuity designs: Extensions*. Cambridge University Press, 2024.

J. McCrary. Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, 142(2):698 – 714, 2008. doi: http://dx.doi.org/10.1016/j.jeconom.2007.05.005.