Causal Inference

Table of contents

1 Overview **2**

2 Potential Outcomes Framework **3**
- 2.1 Potential outcomes . . . 4
- 2.2 Multiple units . . . 5
- 2.3 Assignment mechanism . . . 5
- 2.4 Causal estimands . . . 6
- 2.5 Heterogeneity . . . 8

3 Randomized Experiments **9**
- 3.1 Selection . . . 9
- 3.2 Identification . . . 9
- 3.3 Efficiency . . . 10
- 3.4 Stratification . . . 10
- 3.5 Propensity score . . . 11

4 Instrumental Variables **12**
- 4.1 Compliance . . . 13
- 4.2 Identification . . . 14
- 4.3 Estimation . . . 16
- 4.4 Reduced form . . . 16

5 Observational Studies **17**
- 5.1 Common support . . . 17
- 5.2 Matching . . . 18
- 5.3 Regression . . . 18

6 Difference-in-Differences **19**
- 6.1 2-group-2-period . . . 20
- 6.2 Parallel trends . . . 21

6.3 Regression 21
6.4 2-group-multi-period 22
6.5 Further reading 24

References **24**

1 Overview

In this note we will review some more contemporary topics from the field of Microeconometrics. These are the literatures related to Causal Inference, Treatment Effects, and Policy Evaluation. We will not be able to cover everything. In this handout we will discuss:

1. The Potential Outcomes Framework
2. Randomized Experiments
3. Instrumental Variables
4. Observational Studies
5. Difference-in-Differences

Topics that we will not be able to cover include:

1. Staggered-DiD or Event-studies
2. Synthetic Control
3. Regression Discontinuity Designs

Further reading can be found in:

- Chapters 2.7, 25.1-25.3, 25.5, 25.8 of Cameron and Trivedi (2005)
- Section 7.7 of Verbeek (2017)

Some other texts on the topic include:

- **(MHE)** Angrist and Pischke (2009) *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.
- Angrist (2014) *Mastering Metrics: The Path from Cause to Effect*. Princeton, NJ: Princeton University Press.
- **(IR)** Imbens and Rubin (2015) *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge, UK: Cambridge University Press
- Cunningham (2021) *Causal inference: The Mixtape*. Yale University Press.

MHE is more technical than *Matering 'Metrics*, but with the foundation provided in this module, you should get through much of the content. Note, it was published in 2009 and does not include some more recent topics.

Here are a few papers worth reading:

- Angrist and Krueger (2001)
- Heckman (2008)
- Imbens and Wooldridge (2009)
- Angrist and Pischke (2010)
- Athey and Imbens (2017)

2 Potential Outcomes Framework

The notion of a 'causal effect' is bound by framework within which you define causal effect. In this handout we will build upon the Potential Outcomes (PO) Framework attributed to Jerzy Neyman and Donald Rubin. There are other frameworks, including traditional Econometrics, that seek to model causal relationships (see Heckman and Pinto 2024). Recently, a lot of consideration has been given to the Dynamic Acyclic Graphs approach popularized by Judea Pearl (see Pearl and Mackenzie 2018; also Imbens 2020; Heckman and Pinto 2024). For those interested in the broader question of causality within Economics, take a look at Hoover (2008).

The following is an adaption of the paracetomal example in Imbens and Rubin (2015, 5):

Example 2.1. Consider a new hypothetical policy in the UK, wherein time spent studying in the UK counted towards permanent residency ('indefinite leave to remain'). How would this affect your decision to apply for a graduate visa upon graduation?

There are four potential scenarios:

1. Apply under new rule only:

$$Y(\text{reform}) = \text{apply} \qquad \textit{and} \qquad Y(\text{status quo}) = \text{don't apply}$$

2. Don't apply in either case:

$$Y(\text{reform}) = \text{don't apply} \qquad \textit{and} \qquad Y(\text{status quo}) = \text{don't apply}$$

3. Apply regardless:

$$Y(\text{reform}) = \text{apply} \qquad \textit{and} \qquad Y(\text{status quo}) = \text{apply}$$

4. Don't apply with new rule, but under status quo:

$$Y(\text{reform}) = \text{don't apply} \qquad \textit{and} \qquad Y(\text{status quo}) = \text{apply}$$

We say that in scenarios (2) and (3) there is no causal effect, while in scnearios (1) and (4) there is a causal effect.

Notice two things about this definition of causal effect:

- "First, the definition of the causal effect depends on the potential outcomes, but it does not depend on which outcome is actually observed." (Imbens and Rubin 2015, 6)
- "Second, the causal effect is the comparison of potential outcomes, for the same unit, at the same moment in time post-treatment. In particular, the causal effect is not defined in terms of comparisons of outcomes at different times." (Imbens and Rubin 2015, 6)

There are three important components to the Potential Outcomes framework:

1. **Potential outcomes:** the outcomes corresponding to different levels of treatment or manipulation: "no causation without manipulation" (Rubin 1975, 238).
2. **Multiple units:** given the limitation of observation, one must observe multiple units to infer a causal effect.
3. **Assignment mechanism:** what determines treatment?

2.1 Potential outcomes

For a discrete treatment, we denote the PO of outcome Y as,[1]

$$\begin{aligned} Y_i(1) \quad &\text{if} \quad D_i = 1 \\ Y_i(0) \quad &\text{if} \quad D_i = 0 \end{aligned}$$

where $D_i = \mathbf{1}\{\text{treated}\}$.[2] The observed outcome, Y_i^{obs}, can written as

$$\begin{aligned} Y_i^{obs} = Y_i(D_i) &= Y_i(0) + D_i \cdot (Y_i(1) - Y_i(0)) \\ &= (1 - D_i)Y_i(0) + D_i Y_i(1) \end{aligned}$$

[1] Here, I follow the notation of Guido W. Imbens. You should be comfortable with alternative notations, such as superscripts (Y_i^1, Y_i^0) and subscripts (Y_{i1}, Y_{i0}; as in MM and MHE).

[2] This statement can be generalized to continuous treatment, but it will require some assumptions regarding the causal relationship.

2.2 Multiple units

We cannot observe both POs for any unit i (in the same period of time). At a fundamental level, this means that we cannot observe the unit-level treatment effect:

$$\tau_i = Y_i(1) - Y_i(0)$$

We always need to learn about $f_{(1)}$ and $f_{(0)}$ - the marginal distributions of each potential outcome - from two (or more) samples; be this,

- different units at the same time;
- the same units at different times;
- or a combination of both.

This requires us to make an important assumption: the Stable Unit Treatment Value Assumption (SUTVA).

Definition 2.1. - SUTVA "The potential outcomes for any unit do not vary with the treatments assigned to other units, and, for each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes." (Imbens and Rubin 2015, 10)

This assumption does two things. First, it rules out interference between treatment units. This rules out any spillover effects between 'treatment' and 'control'. Second, it ensures that there are no hidden variations in the treatment level. It is okay for there to be different levels of treatment, so long as they are explicit and well-defined *a priori*.

General equilibrium effects violate SUTVA. In almost all case, we need to assume a partial equilibrium setting. As Economists, this means that the PO framework cannot be used to answer all important empirical questions. There remains a value to more structural models that can identify general equilibrium effects.

2.3 Assignment mechanism

In experimental data, the assignment mechanism is randomization, while in observational data the assignment mechanism is not known.

Randomization is an assignment mechanism that ensures independence/unconfoundedness between the potental outcomes and treatment assignment. This can either be unconditional,[3]

[3]Here, I am using super-population notation. This means that each i is thought of as an *independent* draw from a potentially infinite super-population. For finite sample analysis, the vectors of potential outcomes are treated as non-random. It is only the assignment vector, D, that is random. You would therefore define unconfoundedness as independence across the full vectors of potential outcomes: $Y(1), Y(0) \perp D|X$.

$$Y_i(1), Y_i(0) \perp D_i$$

or conditional on a set of known covariates,

$$Y_i(1), Y_i(0) \perp D_i | X_i$$

The latter is also referred to as the Conditional Independence Assumption (as in MHE & MM) and (strong) ignorability. Both assumptions are referred to as unconfoundedness in the literature.

Note, "[w]ithout unconfoundedness, there is no general approach to estimating treatment effects." (Imbens and Wooldridge 2009, 7). Randomization gives you unconfoundedness, but unconfoundedness can be assumed without randomization; for example, in observational studies. In such cases, you might say it is *as if* assignment is randomized (conditional on X).

Note, in our study of linear models, we made conditional *mean* independence assumptions of the form $E[\varepsilon_i|X_i] = 0$. Mean independence is sufficient for identification in linear models; however, randomization gives you independence between treatment assignment and potential outcomes. You can therefore identify the the marginal distributions $f_{(0)}$ and $f_{(1)}$ and not just their mean $E[Y(0)]$ and $E[Y(1)]$.

2.4 Causal estimands

We cannot identify the unit-level treatment effect, even with randomization. For this reason, the literature focuses on identifying particular causal estimands.

Using the linearity of expectation function, we can identify (from different samples) the Average Treatment Effect from the difference in *uncondtional* means:

$$E[Y_i(1)] - E[Y_i(0)] = E[Y_i(1) - Y_i(0)] = ATE$$

Similarly, we can identify the Average Treatment Effect of the Treated (ATT) and the Untreated (ATU):

$$\begin{aligned} E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 1] =& E[Y_i(1) - Y_i(0)|D_i = 1] = ATT \\ E[Y_i(1)|D_i = 0] - E[Y_i(0)|D_i = 0] =& E[Y_i(1) - Y_i(0)|D_i = 0] = ATU \end{aligned}$$

The above statements are non-trivial. We can learn about the mean of the unit-level treatment effect from different samples. This follows from the fact that the expectation (or average) operator is linear. Unfortunately, this DOES NOT extend beyond the mean.

The q-th percentile of the distribution of unit-level treatment effects, cannot *necessarily* be written as the difference between q-th percentile of $f_{(0)}$ and $f_{(1)}$.

$$F^{-1}_{(1)-(0)}(q) \neq F^{-1}_{(1)}(q) - F^{-1}_{(0)}(q)$$

For the above to be an equality, there must be perfect rank correlation between the distributions of $Y(1)$ and $Y(0)$.

There are additional causal estimands important within this literature:

- Conditional ATE, ATT, and ATU; e.g.,

$$ATE(X_i) = E[Y_i(1) - Y_i(0)|X_i]$$

- Average Causal Effect, typically used to describe continuous (or mutliple dosage) treatments; e.g.,

$$ACE(s) = E[Y_i(s+1) - Y_i(s)|S_i = s]$$

- Local Average Treatment Effects (LATE), as identified by instrumenal variables and regression discontinuity designs.
- Cohort-specific ATT, the ATT for a specific treatment cohort in a staggered difference-in-differences (or event-study); e.g., as denoted by Sun and Abraham (2021)

$$CATT(c) = E[Y_i(c) - Y_i(\infty)]$$

where $Y_i(\infty)$ denotes the PO when "never-treated".

I have ignored an important distinction between finite-sample and super-population estimands. The distinction comes down to whether you think of the sample as the population - in which case, the vectors $Y(1)$ and $Y(0)$ are non-random vectors - or whether you think of the sample as a random draw from a (infinite) super-population. The distinction has important implications for how you compute the variance of an estimator (see Imbens and Rubin 2015, chap. 6). Finite-sample estimands are typically written as averages; hence, the name 'average' TEs. For exaemple, the finite-sample ATE is given by,

$$ATE^{fs} = \frac{1}{N}\sum_{i=1}^{N} Y_i(1) - Y_i(0)$$

The above is an estimand and cannot be computed as it requires you to observe the unit-level treatment effect for all i.

2.5 Heterogeneity

It is important to recognise that treatment effects need not be homogeneous.

Definition 2.2. - Homogenous TEs

$$Y_i(1) - Y_i(0) = \tau \qquad \forall\ i = 1, \dots, n$$

Indeed, you could argue that this literature is largely focused on the estimation of heterogeneous treatment effects. Homogenous treatment effects negate the need to speak about the ATE, ATT, and ATU as all estimands are equivalent and equal to the unit-level TE.

Heterogeneity also has important implications for the use of models to estimate TEs. Consider, if TEs are homogenous, we can write the observed outcome as,

$$\begin{aligned} Y_i =& Y_i(0) + \tau D_i \\ =& E[Y_i(0)] + \tau D_i + Y_i(0) - E[Y_i(0)] \\ =& \alpha + \tau D_i + \varepsilon_i \end{aligned}$$

We can express the observed outcome as a model that is linear in parameters with an error term that is homoskedastic (assuming $Y_i(0)$ is drawn from the same distribution).

With heterogeneous TEs, we get

$$\begin{aligned} Y_i =& Y_i(0) + \tau_i D_i \\ =& E[Y_i(0)] + \tau_{ATE} D_i + Y_i(0) - E[Y_i(0)] + (\tau_i - \tau_{ATE}) D_i \\ =& \alpha + \tau_{ATE} D_i + \upsilon_i \end{aligned}$$

The model only explains the average difference between the treated and control. The heterogeneity remains in the error term. With random assignment, the error term remains conditionally mean independent; however, it is no longer homoskedastic (see Deaton 2010).

In models that include a time dimension - e.g., difference-in-differences - you also need to cosider whether TEs are static. ::: {#def-staticTE} **- Static TEs**

$$Y_{it}(1) - Y_{it}(0) = \tau_i \qquad \forall\ t = 1, \dots, T$$

::: Models with dynamics, will often normalize time relative to the period of treatment; sometimes referred to as event-time. For example, event-time $s = -1$ is the period just before treatment and $s = 0$ the period of treatment.

3 Randomized Experiments

In Economics, randomized experiments are typically referred to as Randomized Control Trials (RCTs); a name borrowed from the Medical field. However, randomization is also used in lab experiments and does sometimes appear in real-world policies (see Angrist 1990).

3.1 Selection

When comparing two groups, the difference between the mean of their outcomes can be decomposed into two terms: the ATT and a selection term.

$$\begin{aligned}
&E[Y_i|D_i = 1] - E[Y_i|D_i = 0] \\
=&E[Y_i(1)|D_i = 1] - [Y_i(0)|D_i = 1] + E[Y_i(0)|D_i = 1] - E[Y_i(0)|D_i = 0] \\
=&E[Y_i(1) - Y_i(0)|D_i = 1] + E[Y_i(0)|D_i = 1] - E[Y_i(0)|D_i = 0] \\
=&ATT + \text{selection}
\end{aligned}$$

where $E[Y_i(0)|D_i = 1]$ is an unobserved counterfactual: the mean of the treated group had it not been treated.

Note, this is a particular definition of selection. When discussing the "problem of selection" in this literature, the question is whether:

$$E[Y_i(0)|D_i = 1] - E[Y_i(0)|D_i = 0] \stackrel{?}{=} 0$$

3.2 Identification

Recall, under randomization assignment is unconfounded; therefore, the observed difference between the mean of the treatment and control group identifies,

$$\begin{aligned}
&E[Y_i|D_i = 1] - E[Y_i|D_i = 0] \\
=&E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 0] \\
=&E[Y_i(1)] - E[Y_i(0)] \\
=&ATE
\end{aligned}$$

where line 3 follows by unconfoundedness. This implies:

1. the selection term = 0;
2. the $ATT = ATE = ATU$.

An RCT can tell you about the ATE of the treatment, for the sampled population. It cannot tell you about the ATE for any unsampled population; an issue referred to as *external validity.*

3.3 Efficiency

One limitation of RCTs is their limited (sample) size. They can be expensive to run, thereby reducing the size of the treated and control sample. For this reason, researchers will take action to design the most efficient (powerful) experiment.

If N (the sample size) is fixed, it is optimal (from an efficiency perspective) to assign $N_t = 0.5N$ units to treatment.

A second action taken by researchers is to control for characteristics in a model; for example,

$$Y_i = \alpha + \beta D_i + X_i'\gamma + \upsilon_i$$

In this instance, these regressors are not included for identification as treatment is unconfounded (this is not the case for stratified experiments; see below). Instead, they are there to reduce the residual variation of the error term.

It is important that these are **good controls**. The included covariates must not be potential outcomes of an experiment. In RCTs, you will typically see that researchers include covariates from a baseline survey.

The OLS estimator for β from the above model is consistent, but may be biased in small-samples depending on whether the heterogeneity of the TE relates to the included X's (Deaton 2010).

$$p\lim \hat{\beta} = ATE$$

3.4 Stratification

Many real-world RCTs stratify assignment into treatment. They divide the sample into blocks - typically based on observable characteristics - and assign each block into treatment independently. By implication, each group can be assigned into treatment with a different probability.

Stratification has an important advantage: since assignmet is unconfounded within each block (or conditional on the stratification covariates X_i), you can conduct more efficient sub-sample analyses.

However, stratification also complicates the use of models to estimate TEs. Since, treatment is independent conditional on X_i, you need to account for X_i in the model.

Suppose, the statification variable(s) takes on m distinct values: $X_i \in \{x_1, x_2, \dots, x_m\}$. Then, the unconditional ATE can be written as a probablity weighted sum of conditional ATEs.

$$ATE = \sum_{j=1}^{m} ATE(x_j) \cdot Pr(X_i = x_j)$$

If you estimate a model, that conditions on each value of X_i (referred to as 'saturated controls'), the coefficient on the treatment indicator gives you a variance weighted average.

$$Y_i = \beta D_i + \sum_{j=1}^{m} \gamma_j \mathbf{1}\{X_i = x_j\} + \zeta_i$$

then,

$$\beta = \frac{\sum_{j=1}^{m} ATE(x_j) Var(W_i | X_i = x_j) Pr(X_i = x_j)}{\sum_{l=1}^{m} Var(W_i | X_i = x_l) Pr(X_i = x_l)}$$

where,

$$Var(D_i | X_i) = Pr(D_i = 1 | X_i) \cdot (1 - Pr(D_i = 1 | X_i))$$

Assigning treatment with different probabilities (based on X_i), implies that the OLS estimator of a linear model with covariates will estimate a variance weighted estimand. If the ATE is independent of X_i, this not a concern, since the weights still sum to 1.

3.5 Propensity score

In this literature, the probability of treatment (conditional on X_i) is referred to as the propensity score:

$$\rho(X_i) = Pr(D_i = 1 | X_i)$$

You can show that the unconditional ATE can be written as an (inverse) propensity weighted estimand:

$$E[Y_i(1)] = E\left[\frac{Y_i \cdot W_i}{\rho(X_i)}\right]$$

and

$$E[Y_i(0)] = E\left[\frac{Y_i \cdot (1 - D_i)}{1 - \rho(X_i)}\right]$$

Which means that,

$$E[Y_i(1) - Y_i(0)] = E\left[\frac{Y_i \cdot D_i}{\rho(X_i)} - \frac{Y_i \cdot (1 - D_i)}{1 - \rho(X_i)}\right] = E\left[\frac{Y_i \cdot (D_i - \rho(X_i))}{\rho(X_i) \cdot (1 - \rho(X_i))}\right]$$

This Weighted Least Squares estimator for β from the univariate model,

$$Y_i = \alpha + \beta D_i + \varepsilon_i$$

using weights,

$$\lambda_i = \frac{1}{\rho(X_i)^{D_i} \cdot (1 - \rho(X_i))^{1 - D_i}}$$

is an unbiased estimator for the unconditional ATE under stratified assignment. This referred to as the Horovitz-Thompson estimator.

This points to a wider issue within this literature: this approach to causal inference is built upon an experiment framework, which does not *necessarily* map to a (linear) regression-model econometrics framework.

4 Instrumental Variables

> **Note**
>
> The following might be a slightly different presentation of instrumental variables to what you are used to from your undergraduate degree.

In some experiments, compliance - i.e. take-up of the treatment - is not a given. This could be for a number of reasons, including ethical reasons. In this case, assignment into the treatment-group does not gaurantee that the individual is treated.

Consider the RCT in Banerjee et al. (2015). The authors randomly assigned a new micro-loan agency to a sub-sample of villages in India. It would be unethical to force a household to take on credit. Instead, the treatment randomly increased *access to* credit. In this application, the mean difference in Y_i between the treatment and control villages cannot identify the causal impact of having a micro-loan. It identifies the ATE of living in a village with increased access to micro-loans.

In imperfect-compliance setting, we adopt the language of Intention to Treat (ITT) effects. In the words of the authors: "given the sampling frame, ours will be an intent-to-treat (ITT) analysis on a sample of "likely borrowers." This is thus neither the effect on those who borrow nor the average effect on the neighborhood. Rather, it is the average effect of easier access to microfinance on those who are its primary targets." (Banerjee et al. 2015, 35).

You might be wondering, why can't we just condition on receipt of the treatment? If compliance is imperfect, then take-up of the treatment is endogenous: determined by factors other than random assignment. As a result the treatment-receiving sample is no longer randomly assigned.

You can have both one-sided and two-sided non-compliance. Here, we will focus on the latter, as it more closely relates to the use of instruments in Applied Economics (for example Angrist 1990). One-sided non-compliance means that some individuals in the treated group may not receive the treatment, but does not allow for the case where individuals in the control group may receive it. Two-sided non-compliance allows for both.

4.1 Compliance

Let $D_i \in \{0, 1\}$ denote receipt of treatment and let $Z_i \in \{0, 1\}$ denote treat-group assignment. With two-sided non-compliance, we can consider the potential outcomes of D_i:

$$D_i(Z_i) \in \{D_i(1), D_i(0)\}$$

Using these potential outcomes we can assign names to each possible compliance status:

Table 1: Compliance status:

	$D_i(1) = 0$	$D_i(1) = 1$
$D_i(0) = 0$	never-treated (nt)	complier (c)
$D_i(0) = 1$	defier (d)	always-treated (at)

Define $G_i \in \{nt, c, d, at\}$, with $Pr(G_i = g) \in \{\rho_{nt}, \rho_c, \rho_d, \rho_{at}\}$.

When you look at the data, you don't observe each potential outcome. Instead, you observe the pair $\{Z_i, D_i\}$; each combination of which will contain a mix of the above groups.

Table 2: Compliance status:

	$D_i = 0$	$D_i = 1$
$Z_i = 0$	never-treated + compliers	always-treated + defiers
$Z_i = 1$	never-treated + defiers	always-treated + compliers

The ITT of assignment on treatment receipt is given by:

$$\begin{aligned} ITT_D =& E[D_i(1) - D_i(0)] \\ =& E[D_i(1)] - E[D_i(0)] \\ =& \rho_c + \rho_{at} - (\rho_d + \rho_{at}) \\ =& \rho_c - \rho_d \end{aligned}$$

This is the difference in the proportion of compliers (those who take-up treatment only when assigned) and defiers (those who take up treatment only when not assigned) in the (super) population.

The potential outcomes for the outcome variable are:

$$Y_i(Z_i, D_i(Z_i)) \in \{Y_i(0,0), Y_i(0,1), Y_i(1,0), Y_i(1,1)\}$$

For example, $Y_i(1,0)$ is the potential outcome of unit i were they to receive assignment into treatment, but not take it up. Likewise, $Y_i(1,1)$ represents the potential outcome of unit i were they to take-up the treatment after being assigned it.

The ITT of assignment on the main outcome is:

$$ITT_Y = E[Y_i(1) - Y_i(0)] = E[Y_i(1, D_i(1)) - Y_i(0, D_i(0))]$$

This term is what you identify when comparing the means of the outcome in the treated and control groups, under randomization.

Under randomization, the instrument is unconfounded with respect to both sets of potential outcomes:

$$Z_i \perp D_i(1), D_i(0), Y_i(0,0), Y_i(0,1), Y_i(1,0), Y_i(1,1)$$

4.2 Identification

In addition to unconfoundedness, we require the following exclusion restrictions assumptions:

- for all never-takers: $Y_i(1,0) = Y_i(0,0)$
- for all always-takers: $Y_i(0,1) = Y_i(1,1)$

Then we need a final **monotonicity** (i.e., no defiers) assumption:

- no defiers: $D_i(1) \geq D_i(0) \Rightarrow \rho_d = 0$

Under these four assumptions:

$$\frac{ITT_Y}{ITT_D} = E[Y_i(1) - Y_i(0)|G_i = c] = LATE$$

This is referred to as the Local Average Treatment Effect: the ATE of compliers. Note, you cannot directly observe the population of compliers in the data, as this would require observing both states of the outcome $D_i(Z_i)$. In contrast, when we identify the ATT (as with DiD), this the ATE of the observed Treated population.

We can prove this result as follows. In the denominator, we have:

$$\begin{aligned} E[D_i(1) - D_i(0)] =& \rho_{nt} E[D_i(1) - D_i(0)|G_i = nt] \\ &+ \rho_{at} E[D_i(1) - D_i(0)|G_i = at] \\ &+ \rho_c E[D_i(1) - D_i(0)|G_i = c] \\ =& \rho_c \end{aligned}$$

since by the monotonicity (no defiers) assumption $\rho_d = 0$; $D_i(1) = D_i(0)$ for never- and always-takers; and $D_i(1) - D_i(0) = 1$ for compliers.

In the numerator, we have:

$$\begin{aligned} & E[Y_i(1, D_i(1)) - Y_i(0, D_i(0))] \\ =& E[Y_i(1,1)|D_i(1) = 1] \cdot Pr(D_i(1) = 1) + E[Y_i(1,0)|D_i(1) = 0] \cdot Pr(D_i(1) = 0) \\ & - E[Y_i(0,1)|D_i(0) = 1] \cdot Pr(D_i(0) = 1) - E[Y_i(0,0)|D_i(0) = 0] \cdot Pr(D_i(0) = 0) \end{aligned}$$

Assuming no defiers,

$$\begin{aligned} E[Y_i(0,1)|D_i(0) = 1] \cdot Pr(D_i(0) = 1) =& \rho_{at} E[Y_i(0,1)|G_i = at] \\ =& \rho_{at} E[Y_i(1,1)|G_i = at] \end{aligned}$$

Where the last line follows from the exclusion restriction. Similarly,

$$\begin{aligned} E[Y_i(1,0)|D_i(1) = 0] \cdot Pr(D_i(1) = 0) =& \rho_{nt} E[Y_i(1,0)|G_i = nt] \\ =& \rho_{nt} E[Y_i(0,0)|G_i = nt] \end{aligned}$$

Applying the exclusion restrictions we get:

$$\begin{aligned} =& \rho_c E[Y_i(1,1)|G_i = c] - \rho_c E[Y_i(0,0)|G_i = c] \\ & + \rho_{at} E[Y_i(1,1)|G_i = at] - \rho_{at} E[Y_i(1,1)|G_i = at] \\ & + \rho_{nt} E[Y_i(0,0)|G_i = nt] - \rho_{nt} E[Y_i(0,0)|G_i = nt] \\ =& \rho_c E[Y_i(1,1) - Y_i(0,0)|G_i = c] \end{aligned}$$

Dividing by the ITT_D, we get the LATE result.

4.3 Estimation

The LATE can be estimated by two-stage-least-squares (2SLS).

1. First, estimate the first-stage (ITT_D):

$$D = \phi_1 + \phi_2 Z + \nu$$

2. Second, estimate the second-stage, substituting D with the predicted values of the first-stage:[4]

$$Y = \beta_1 + \beta_2 P_Z D + \epsilon$$

In this just-identified case,

$$\hat{\beta}_2^{2SLS} = \frac{\hat{\gamma}_2}{\hat{\phi}_2}$$

where $\hat{\gamma}_2$ is the OLS estimator of the reduced form equation,

$$Y = \gamma_1 + \gamma_2 Z + \zeta$$

Recall, the reduced form equation identifies the ITT_Y.

4.4 Reduced form

Recall, Z denotes random treatment-assignment. Much of empirical in applied microeconomics takes on this form: regressing the outcome directly on treatment-assignment. It is for this reason that it is often referred to as "reduced form" research.

However, the phrase "reduced form" comes from an older literature, where the regression of Y on D, where D was potentially endogenous, was a "structural equation", including parameters from a structural model. In this setting, we do not typically adopt this phrase to describe the relationship between Y and D.

This framework provides a useful perspective on experiments. Instruments are central to the way economists think about causation (see Angrist and Krueger 2001), and randomized experiments provide the ideal instrumental variable. You can either examine the reduced-form relationship, given by treatment-assignment, or use the experiment as an instrument to study an endogenous relationship.

[4] This is equivalent to including the control function: $Y = \beta_1 + \beta_2 Z + \beta_3 M_Z D + \varepsilon$.

Consider the following setting. Suppose, you are interested in the causal relationship between household income (RHS) and investment into child-education (LHS) in poorer countries. Evidently household income (a continuous variable) is endogenous, and comparing households with high and low incomes will give rise to differences largely explained by selection.

We need a source of exogenous variation in household income. We could design a RCT that enrolls a random sample of households into a universal basic income (UBI) scheme. Comparing the difference in education of households in treatment and control, would give us the reduced form effect of the experiment. Alternatively, we could use treatment assignment as an instrument for household income. Of course, we hope to find that treatment increases household income by an amount close to the value of the UBI scheme.

5 Observational Studies

In observational studies, the functional form of the assignment mechanism is not known. The standard approach in these settings is to assume Conditional Independence (CIA/unconfoundedness). Afterall, "[w]ithout unconfoundedness, there is no general approach to estimating treatment effects" (Imbens and Wooldridge, 2009, p.7).

$$Y_i(1), Y_i(0) \perp D_i | X_i$$

This assumption suggests that *conditional on* X it is *as if* assignment is randomized. Thus,

$$E[Y_i(0)|D_i = 1, X_i] - E[Y_i(0)|D_i = 0, X_i] = 0$$

There is no selection, conditional on X_i. The above statement, is also referred to as "selection on observables". The treatment and control group can differ in terms of the distribution of X_i; however, conditional on X_i they are balanced. This rules out "selection on unobservables".

5.1 Common support

In controlled experiments the researcher picks the number of treated and control units. There is just one requirement: $1 \leq N_t \leq N-1$, at least one unit must be (un)treated. In observational studies, you do not have direct control over the assignment mechanism. We therefore need an additional assumption: **common support** (or overlap).

Definition 5.1. - Common Support

$$0 < \rho(X_i) < 1$$

This assumption ensures that for every X_i, you observe both treated and control units with a non-zero probability.

5.2 Matching

The most direct approach to matching is covariate-matching: given common support, you can estimate a Conditional ATE for each X_i, after which you aggregate up to the unconditional ATE using the distribution of X.

$$ATE = \sum_{j=1}^{m} ATE(x_j) \cdot Pr(X_i = x_j)$$

Unfortunately, covariate matching quickly encounters the **curse of dimensionality**. You need to match on all possible values of the vector $X_i \in \mathbb{R}^k$. Consider, if each of the X_{ik} covariates is a dummy variable, then the number of possible values for the vector X_i is 2^k.

In addition, there is the challenge of matching on continuous variables. This can be solved by discretizing the support of the continuous variable, but necessarily increases the number of values to matching on.

An alternative is Propensity Score Matching (PSM). Rosenbaum and Rubin (1983) show that you can match on the propensity score instead of covariates. This is because the propensity score is a balancing score.

Definition 5.2. - Balancing Score: A function $b : \mathbb{R}^k \to \mathbb{R}$, such that,

$$D_i \perp X_i | b(X_i)$$

You can show that $\rho(X_i) = Pr(D_i = 1|X_i)$ is a balancing score, by demonstrating the equality between:

$$Pr(D_i = 1|X_i, \rho(X_i)) = Pr(D_i = 1|\rho(X_i))$$

Given this result, unconfoundedness implies unconfoundedness *on the propensity score*,

$$Y_i(1), Y_i(0) \perp D_i | \rho(X_i)$$

Take a look at Hirano, Imbens, and Ridder (2003) and Caliendo and Kopeinig (2008), if you are interested in some of the practicialities of PSM estimation. The major challenge for PSM, in observational study settings, is that we do not know the true functional form of $\rho(X_i)$.

5.3 Regression

You can think of linear regression as a matching estimator. Recall, we can always write:

$$Y_i = E[Y_i|D_i, X_i] + \varepsilon_i$$

where $E[\varepsilon_i|D_i, X_i] = 0$. Given that Y_i is a combination of potential outcomes,

$$Y_i = E[Y_i(0)|D_i, X_i] + D_i E[Y_i(1) - Y_i(0)|D_i, X_i] + \varepsilon_i$$

Assuming CIA/unconfoundedness,

$$Y_i = E[Y_i(0)|X_i] + D_i E[Y_i(1) - Y_i(0)|X_i] + \varepsilon_i$$

If we are willing to assuming that $E[Y_i(0)|X_i]$ is linear (in parameters), then we get

$$Y_i = X_i'\gamma + D_i E[Y_i(1) - Y_i(0)|X_i] + \varepsilon_i$$

We then need to make an assumption concerning the heterogeneity w.r.t. X. Alternatively, we can opt for a more flexible saturated-controls model,

$$Y_i = \sum_{j=1}^{m} \gamma_j \mathbf{1}\{X_i = x_j\} + D_i E[Y_i(1) - Y_i(0)|X_i] + \zeta_i$$

Clearly, we need to make an assumption concerning how heteroegeneity relates to X_i. If it is independent of X_i, then the saturated model will approximate any functional form of $E[Y_i(0)|X_i]$, but is limited by the curse of dimensionality.

Suppose we are unwilling to assume homogeneity w.r.t. X. If we allow the coefficient on D to vary with every value of X, we are back to covariate-matching (i.e. a separate model for each covariate value). If we estimate a single parameter coefficient for D, we end up with variance weighting. There is no obvious path forward.

We also need to take note of common-support. Saturated-controls models require common-support, but linear models do not. You should be aware that linear models can allow you to extrapolate a counterfactual to parts of the data without common support.

6 Difference-in-Differences

Difference-in-differences (DiD) is one of the most commonly used empirical strategies in applied microeconomics research. This is partly because the method is well suited to the study of "natural"/quasi-experiments. These are empirical settings where the conditions of an experiment are met - that is, assignment into a treatment and control group - but the researcher has no control over the assignment. For example, a natural disaster that shocks a particular region, but does not affect a neighbouring region (see Card 1990). Or, the introduction of a new policy that affects one group of people, but not another.

Crucially, DiD does *not* require unconfoundedness. Instead, identification is based on a parallel trends assumption, as well as a few exclusion restrictions. The DiD approach utilizes a time-dimesion that has so far been ignored. It also has the advantage of being feasible with repeated cross-sections of data, as panel data is not always avaiable.

6.1 2-group-2-period

The simple 2-group-2-period DiD set-up has the following characteristics,

- treatment takes place in period t_0;
- you observe both treated and control samples in a period before and after treatment;
- and there exists a never-treated control group.

As before, let the time-invariant dummy variable, D_i, denote the treatment-group status of unit i. Next, let the time-varying dummy variable, $T_t = \mathbf{1}\{t \geq t_0\}$, be $= 1$ in the period of treatment (t_0) and 0 before.

We will make the following exclusion restrictions,

Definition 6.1. - Exclusion Restriction: No Anticipation (no pre-emptive behaviour)

$$Y_{it} = Y_{it}(1) = Y_i t(0) \qquad \forall\ (i,t) \text{ s.t. } \ t < t_0$$

Some texts will assume random/unexpected timing of the treatment, as a mechanism that rules out anticipation. Note, this assumption is stronger than what is required for identification of the ATT. For example, Wooldridge (2023) assumes the a weaker assumption:

$$E[Y_{it}(1)|D_i = 1, T_t = 0] = E[Y_{it}(0)|D_i = 1, T_t = 0]$$

Definition 6.2. - Exclusion Restriction: No Spillovers

$$Y_{it} = Y_{it}(0) \qquad \forall\ (i,t) \text{ s.t. } D_i = 0$$

Evidently, this assumption is met by SUTVA.

Then, we can write the observed Y_{it} as,

$$\begin{aligned} Y_{it} = &\begin{cases} Y_{it}(0) & \forall \quad t < t_0 \\ Y_{it}(0) + D_i \cdot (Y_{it}(1) - Y_{it}(0)) & \forall \quad t \geq t_0 \end{cases} \\ =&Y_{it}(0) + T_t \cdot D_i \cdot (Y_{it}(1) - Y_{it}(0)) \end{aligned}$$

6.2 Parallel trends

The main identifying assumption of a DiD model is parallel trends. This can be stated as,

Definition 6.3. - Parallel Trends

$$\begin{aligned} &E[Y_{it}(0)|D_i = 0, T_t = 1] - E[Y_{it}(0)|D_i = 0, T_t = 0] \\ =&E[Y_{it}(0)|D_i = 1, T_t = 1] - E[Y_{it}(0)|D_i = 1, T_t = 0] \end{aligned}$$

Under these assumptions, the difference-in-differences (DiD) of conditional means gives you the ATT.

$$\begin{aligned} &\big[E[Y_{it}|D_i = 1, T_t = 1] - E[Y_{it}|D_i = 1, T_t = 0]\big] \\ &- \big[E[Y_{it}|D_i = 0, T_t = 1] - E[Y_{it}|D_i = 0, T_t = 0]\big] \\ =&\big[E[Y_{it}(1)|D_i = 1, T_t = 1] - E[Y_{it}(0)|D_i = 1, T_t = 0]\big] \\ &- \big[E[Y_{it}(0)|D_i = 0, T_t = 1] - E[Y_{it}(0)|D_i = 0, T_t = 0]\big] \\ =&\big[E[Y_{it}(1)|D_i = 1, T_t = 1] - E[Y_{it}(0)|D_i = 1, T_t = 1]\big] \\ &+ \big[E[Y_{it}(0)|D_i = 1, T_t = 1] - E[Y_{it}(0)|D_i = 1, T_t = 0]\big] \\ &- \big[E[Y_{it}(0)|D_i = 0, T_t = 1] - E[Y_{it}(0)|D_i = 0, T_t = 0]\big] \\ =&E[Y_{it}(1) - Y_{it}(0)|D_i = 1, T_t = 1] \\ =&ATT(t_0) \end{aligned}$$

Where, parallel trends implies that the difference between the last two parentheses in line 3 is 0. Note, I have denoted this as the ATT in period t_0 as the treatment effect may not be static.

6.3 Regression

This DiD estimand can be expressed within a linear regression model,

$$Y_{it} = \alpha + \psi D_i + \delta T_t + \beta D_i \times T_t + \varepsilon_i$$

where β, the parameter on the interaction term is given by the above DiD.

In the above model, the $E[\varepsilon_i|D_i, T_t] = 0$, since the model is fully saturated. However, this does not mean that the $\beta = ATT$, it just means that β gives you the DiD of conditional means and the OLS estimator, $\widehat{\beta}^{DD}$, is an unbiased estimator. The fact that $\beta = ATT$ depends on the parallel trends assumption, stated in terms of the potential outcome $Y_{it}(0)$. This is a perfect example of how identification of (population) parameters within a regression model is different to identification of TEs.

Given this linear regression model, some texts will state the parallel trends in parametric form as:

Definition 6.4. - Parallel Trends (parametric version)

$$E[Y_{it}(0)|D_i = 0, T_t = 1] = \alpha + \psi D_i + \delta T_t$$

The two definitions are equivalent, if you consider the definition of the parameters in the above expression.

The above specification does not require longitudinal data; however, if it is available you can then estimate the Fixed Effects model:

$$Y_{it} = \alpha_i + \delta T_t + \beta D_i \times T_t + \varepsilon_i$$

Controlling for unit-FEs tends to yield a more efficient estimator, as the unit-level dummies explain more of the variation in Y. Moreover, with a balanced panel, you can show:

$$\hat{\beta}^{LSDV} = \hat{\beta}^{DD}$$

The result arises from the fact that the regressors in the simple DiD model provide the same projection of the interaction term as the FE model which replaces the constant and D-dummy with unit-specific dummy variables.

6.4 2-group-multi-period

We can extend the above set-up to include multiple periods, before and after the treatment. We do then need to add an assumption that the treatment is absorbing (or 'always-on'). We can also now allow for dynamic TEs.

With multiple periods, the above specification (including regressors $[1, D, T, D \times T]$) is referred to as a 'static' specification. This is because it estimates a single ATT for all post-treatment periods; as would be appropriate if the TE is indeed static.

It is preferable to estimate one of the following dynamic specifications:

- semi-dynamic specification

$$Y_{it} = \psi D_i + \delta_t + \sum_{j \geq t_0} \beta_j D_i \times \mathbf{1}\{t = j\} + \varepsilon_i$$

  or with panel data,

$$Y_{it} = \alpha_i + \delta_t + \sum_{j \geq t_0} \beta_j D_i \times \mathbf{1}\{t = j\} + \upsilon_i$$

- fully-dynamic specification

$$Y_{it} = \psi D_i + \delta_t + \sum_{j \neq t_0 - k} \beta_j D_i \times \mathbf{1}\{t = j\} + \varepsilon_i$$

or with panel data,

$$Y_{it} = \alpha_i + \delta_t + \sum_{j \neq t_0 - k} \beta_j D_i \times \mathbf{1}\{t = j\} + \upsilon_i$$

for a chosen base period (in the pre-period): $k \geq 1$.

The fully-dynamic specification is preferred for the following reason. The pre-treatment β_j coefficients ($j < t_0$), provide a valid test for parallel trends **in the pre-period**. This can be used to *support* the assumption of parallel trends in the post-period. Note, this is NOT a test of parallel trends in the post-period when it needs to hold.

Suppose $k = 1$, then the base period is $t_0 - 1$. Assuming no anticipation, we can then write β_{t_0-2} as,

$$\begin{aligned}
\beta_{t_0-2} =& [E[Y_{it}|D_i = 1, t = t_0 - 2] - E[Y_{it}|D_i = 1, t = t_0 - 1]] \\
& - [E[Y_{it}|D_i = 0, t = t_0 - 2] - E[Y_{it}|D_i = 0, t = t_0 - 1]] \\
=& [E[Y_{it}(0)|D_i = 1, t = t_0 - 2] - E[Y_{it}(0)|D_i = 1, t = t_0 - 1]] \\
& - [E[Y_{it}(0)|D_i = 0, t = t_0 - 2] - E[Y_{it}(0)|D_i = 0, t = t_0 - 1]]
\end{aligned}$$

The test, $H_0 : \beta_{t_0-2} = 0$ is a valid test for parallel trends between period $t_0 - 2$ and $t_0 - 1$ (assuming no anticipation).

Suppose $k = 2$, then the base period is $t_0 - 2$. Assuming parallel trends, we can then write β_{t_0-1} as,

$$\begin{aligned}
\beta_{t_0-1} =& [E[Y_{it}|D_i = 1, t = t_0 - 1] - E[Y_{it}|D_i = 1, t = t_0 - 2]] \\
& - [E[Y_{it}|D_i = 0, t = t_0 - 1] - E[Y_{it}|D_i = 0, t = t_0 - 2]] \\
=& [E[Y_{it}(1) - Y_{it}(0)|D_i = 1, t = t_0 - 1] \\
=& ATT(t_0 - 1)
\end{aligned}$$

The test, $H_0 : \beta_{t_0-1} = 0$ is a valid test for no anticipation in period $t_0 - 1$ (assuming parallel trends). Crucially, in both instances, we must assume one assumption to test the other. That is, you cannot disentangle a pre-emptive behaviour from a failure of parallel trends in the data.

6.5 Further reading

There is a large body of literature that discusses a range of topics related to DiD models. Two topics you should take note of when applying these methods are: the appropriate estimation of SEs, and heterogeneity in staggered DiDs (also referred to as event-studies). I have provided a few citations below.

Take a look at the following texts concerning clustered SEs within DiD models:

- Wooldridge (2003)
- Bertrand, Duflo, and Mullainathan (2004)
- Abadie et al. (2023)

Take a look at the following texts to learn more about staggered DiD (event-study) models:

- Imai and Kim (2019)
- Clément De Chaisemartin and D'Haultfœuille (2020); Clement De Chaisemartin and D'Haultfœuille (2023); and Clément De Chaisemartin and D'Haultfœuille (2023)
- Sun and Abraham (2021) (see also Stata package `eventstudyweights` and `eventstudyinteract`)
- Callaway and Sant'Anna (2021) (see also Stata package `csdid` and `drdid`)
- Goodman-Bacon (2021)
- Athey and Imbens (2022)
- Borusyak, Jaravel, and Spiess (2024)

Take a look at the following texts to learn more about parallel trends for non-linear models:

- Wooldridge (2023)

Take a look at the following texts to learn more about conditional parallel trends:

- Caetano and Callaway (2024)

References

Abadie, Alberto, Susan Athey, Guido W Imbens, and Jeffrey M Wooldridge. 2023. "When Should You Adjust Standard Errors for Clustering?" *The Quarterly Journal of Economics* 138 (1): 1–35.

Angrist, Joshua D. 1990. "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records." *The American Economic Review*, 313–36.

———. 2014. *Mastering'metrics: The Path from Cause to Effect*. Princeton University Press.

Angrist, Joshua D, and Alan B Krueger. 2001. "Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments." *Journal of Economic Perspectives* 15 (4): 69–85.

Angrist, Joshua D, and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton university press.

———. 2010. "The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con Out of Econometrics." *Journal of Economic Perspectives* 24 (2): 3–30.

Athey, Susan, and Guido W Imbens. 2017. "The State of Applied Econometrics: Causality and Policy Evaluation." *Journal of Economic Perspectives* 31 (2): 3–32.

———. 2022. "Design-Based Analysis in Difference-in-Differences Settings with Staggered Adoption." *Journal of Econometrics* 226 (1): 62–79.

Banerjee, Abhijit, Esther Duflo, Rachel Glennerster, and Cynthia Kinnan. 2015. "The Miracle of Microfinance? Evidence from a Randomized Evaluation." *American Economic Journal: Applied Economics* 7 (1): 22–53.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2004. "How Much Should We Trust Differences-in-Differences Estimates?" *The Quarterly Journal of Economics* 119 (1): 249–75.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess. 2024. "Revisiting Event-Study Designs: Robust and Efficient Estimation." *Review of Economic Studies*, rdae007.

Caetano, Carolina, and Brantly Callaway. 2024. "Difference-in-Differences When Parallel Trends Holds Conditional on Covariates." *arXiv Preprint arXiv:2406.15288*.

Caliendo, Marco, and Sabine Kopeinig. 2008. "Some Practical Guidance for the Implementation of Propensity Score Matching." *Journal of Economic Surveys* 22 (1): 31–72.

Callaway, Brantly, and Pedro HC Sant'Anna. 2021. "Difference-in-Differences with Multiple Time Periods." *Journal of Econometrics* 225 (2): 200–230.

Cameron, A Colin, and Pravin K Trivedi. 2005. *Microeconometrics: Methods and Applications*. Cambridge university press.

Card, David. 1990. "The Impact of the Mariel Boatlift on the Miami Labor Market." *Ilr Review* 43 (2): 245–57.

Cunningham, Scott. 2021. *Causal Inference: The Mixtape*. Yale university press.

De Chaisemartin, Clement, and Xavier D'Haultfœuille. 2023. "Two-Way Fixed Effects and Differences-in-Differences Estimators with Several Treatments." *Journal of Econometrics* 236 (2): 105480.

De Chaisemartin, Clément, and Xavier D'Haultfœuille. 2020. "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects." *American Economic Review* 110 (9): 2964–96.

———. 2023. "Two-Way Fixed Effects and Differences-in-Differences with Heterogeneous Treatment Effects: A Survey." *The Econometrics Journal* 26 (3): C1–30.

Deaton, Angus. 2010. "Instruments, Randomization, and Learning about Development." *Journal of Economic Literature* 48 (2): 424–55.

Goodman-Bacon, Andrew. 2021. "Difference-in-Differences with Variation in Treatment Timing." *Journal of Econometrics* 225 (2): 254–77.

Heckman, James J. 2008. "Econometric Causality." *International Statistical Review* 76 (1): 1–27.

Heckman, James J, and Rodrigo Pinto. 2024. "Econometric Causality: The Central Role of

Thought Experiments." *Journal of Econometrics*, 105719.

Hirano, Keisuke, Guido W Imbens, and Geert Ridder. 2003. "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score." *Econometrica* 71 (4): 1161–89.

Hoover, Kevin D. 2008. "The New Palgrave Dictionary of Economics." In, edited by Steven N. Durlauf and Lawrence E. Blume, 2nd ed. Palgrave Macmillan.

Imai, Kosuke, and In Song Kim. 2019. "When Should We Use Unit Fixed Effects Regression Models for Causal Inference with Longitudinal Data?" *American Journal of Political Science* 63 (2): 467–90.

Imbens, Guido W. 2020. "Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics." *Journal of Economic Literature* 58 (4): 1129–79.

Imbens, Guido W, and Donald B Rubin. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge university press.

Imbens, Guido W, and Jeffrey M Wooldridge. 2009. "Recent Developments in the Econometrics of Program Evaluation." *Journal of Economic Literature* 47 (1): 5–86.

Pearl, Judea, and Dana Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect*. Basic books.

Rosenbaum, Paul R, and Donald B Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70 (1): 41–55.

Rubin, Donald B. 1975. "Bayesian Inference for Causality: The Importance of Randomization." In *The Proceedings of the Social Statistics Section of the American Statistical Association*, 233:239. American Statistical Association Alexandria, VA.

Sun, Liyang, and Sarah Abraham. 2021. "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects." *Journal of Econometrics* 225 (2): 175–99.

Verbeek, Marno. 2017. *A Guide to Modern Econometrics*. John Wiley & Sons.

Wooldridge, Jeffrey M. 2003. "Cluster-Sample Methods in Applied Econometrics." *American Economic Review* 93 (2): 133–38.

———. 2023. "Simple Approaches to Nonlinear Difference-in-Differences with Panel Data." *The Econometrics Journal* 26 (3): C31–66.