Harvard University
Stat 286: Course Notes

Kyla Chasalow

kyla_chasalow@g.harvard.edu

Fall 2023

Preliminaries

- I created these notes in Fall 2023 to accompany Stat 286 taught by Kosuke Imai at Harvard and TAed by myself and another graduate student. Then otes began as review sheets for my weekly sections, but soon expanded to a fuller treatment of the subjects covered in the course lectures. That said, there are some topics which were covered in 'extension lectures' that are not included here and some topics which are treated only very briefly. My main goal is to summarize many of the most important concepts for each unit and provide some tips on things I imagine might be confusing or was confused by myself when I first encountered them. I hope that these course notes will remain useful for future iterations of the course.
- Imbens and Rubin (2015) is the main text for the course but does not cover all the modules. I have at times listed other resources in Further Resources sections but note that these materials are 100% optional for the course and will not necessary for any problem set or exam!
- In places, I mention that parts of the notes are relevant to certain problem sets or review sheets. These may change in future iterations of the course.
- Pr(there exists at least one typo below) > 0. Please let me know if you find one.

Contents

1 Module 1: Potential Outcomes and Causality

Running Example: Tutoring Experiment: Suppose we run an experiment at a school to see if a tutoring program causes children to pass a math test. We have binary treatment $T = 1$ (tutoring) and $T = 0$ (no tutoring) and binary outcomes $Y = 1$ (pass math test) and $Y = 0$ (does not pass math test).

1.1 Key Concepts

1. **Unit**: The object of analysis (e.g., school children) that is measured in an experiment run in a particular context (e.g., School X, Fall 2023, 3rd grade)

2. **Treatment (intervention, policy)**: an action or condition that is applied to units, must have at least 2 possible values (e.g., tutoring, no tutoring)

3. **Observed Outcome Y_i**: the value of outcome variable Y that is recorded in the data for unit i (e.g., whether student i passes a math test).

4. **Potential Outcome $Y_i(t)$**: the outcome for unit i if unit i receives treatment t in the context of a particular experiment. If unit i actually receives treatment t, then the potential outcome under t and the **observed outcome** are the same $Y_i(t) = Y_i$. If unit i does not receive treatment t, $Y_i(t)$ is an unobservable **counterfactual**: what *would have been observed* if unit i got treatment t.

5. **Fundamental Problem of Causal Inference**: cannot observe multiple potential outcomes for a given unit.

6. **Principle Strata** A partition of units by the values of their potential outcomes under each treatment. Unobservable but relevant to defining causal effects and estimands. See Module 1 problem set on surrogate outcomes for a good example of how they can be useful. Because potential outcomes are viewed as fixed, principal strata membership is a **pre-treatment variable** (as opposed to a **post-treatment** variable that might be affected by treatment).

Stratum $(Y(0), Y(1))$	Unit causal effect	Description
$(1, 1)$	0	Passes math test whether tutored or not
$(0, 1)$	1	Only passes math test if tutored
$(1, 0)$	-1	Only passes math test if not tutored
$(0, 0)$	0	Fails math test whether tutored or not

Table 1: Principal strata for tutoring example

7. **Causality**: refers to differences in potential outcomes across counterfactuals. Saying a treatment causes an outcome implies that the outcome would have been different under other treatment(s). For example, if a student would pass the math test with or without tutoring, there is no causal effect of tutoring on passing the math test for that student.

8. **Unit-Level Causal Effect**: A function of a unit's potential outcomes. If not otherwise specified, this means the additive effect $Y_i(1) - Y_i(0)$ but other functions such as $Y_i(1)/Y_i(0)$ could be of interest.

9. **Notation $\mathcal{O}_n$** In coming modules, we will often use $\mathcal{O}_n = \{Y_i(0), Y_i(1)\}_{i=1}^n$ to denote the set of potential outcomes for a sample of n people. For a given finite set of people, we treat $\mathcal{O}_n$ as fixed (though never fully observable)

1.2 Standard Assumptions

In this course, we almost always assume:

- **Causal ordering**: treatment T causes Y and not the other way around. We assume no simultaneity (T and Y causing each other)
- **Consistency**: $Y_i = Y_i(t)$ whenever $T_i = t$. This says that there is no hidden variation in what it means to receive treatment t that then leads to variation in potential outcomes. This is violated if, for example, I define treatment as $t =$ "receive tutoring once per week," but some children actually received tutoring once per month and $Y_i = Y_i(\text{tutor once per month}) \neq Y_i(\text{tutor once per week})$. In real situations, there is often some variety in what it actually means to receive treatment (here: frequency, teacher quality, time of day, more). One way to resolve this is to redefine treatment e.g., to mean "receive at least some tutoring of some form" and to think of treatment effects in terms of 'on average' effects over variations of the treatment. This does change the nature of what you are studying, and major variation in treatment can still dilute or alter the effect you detect - redefining does not magic away the underlying issue.
- **No Interference**: $Y_i(T_1, ..., T_n) = Y_i(T_i) \quad \forall i \in \{1, ..., n\}$. This says that unit i's outcome is only a function of unit i's treatment. This is violated if child i and child j are friends and when child i is tutored, child j learns from child i and benefits, too. If there is interference, these causal effects are called **spillover effects**.

Note: consistency + no interference are often referred to as **SUTVA** - Stable Unit Treatment Value Assumption.

1.3 Causal Identification

If $Y_i(0)$ is fundamentally unobservable for a person who recieves treatment, why do we care about it? Are principal strata relevant given we'll never know who belongs to which one? A key thing about causal inference that may be different from what you've encountered in other statistics classes is the huge emphasis on the task of **identification**: linking the things we want to know but fundamentally cannot observe to things we can observe and which, *at least in expectation* or as sample size $n \to \infty$, tell us about what we want to know. There are three key steps:

1. **Estimand** - defining what we want to know. This is where causal inference is very strong – it gets to the heart of what we actually want to know in science and policy.
2. **Identification** - can we link what we want to know to something observable so that, at least if we had infinite data, we could learn it? In principle, I can declare I want to know anything. Maybe, just for fun, I want to know the value of $Y_i(0)^2 e^{Y_i(1)} + \cos(5Y_i(1))$ or, more realistically, of $Y_i(1) - Y_i(0)$. Unfortunately, both of these quantities are fundamentally unknowable. They are **non-identifiable**. The *average treatment effect* (to come in Module 3) is identifiable. I will provide some more formal notes on identification in that module.
3. **Estimation** - back to statistics! In reality, we have finite data. We need to develop **estimators** that give us an estimate of what we want to know. We should also think about how to quantify uncertainty in those estimates via variance estimates and confidence intervals. Samples size is a key concern here.

Implicitly, even in your past statistics classes, you made identification assumptions. You've probably assumed, for example, that the data were *i.i.d.* and used this to argue that an estimator $\hat{\theta}$ was unbiased for some true θ. In causal inference, however, we more obviously hit identification issues because of the fundamental problem of causal inference.

1.4 Tips

1. **Potential outcomes are not random for an individual unit.** $Y_i(t)$ is a fixed attribute of unit i. E.g., suppose child i would pass the math test if tutored ($Y_i(1) = 1$) and fail if not tutored ($Y_i(0) = 0$). We think of these things as attributes of the child. One attribute is **revealed** by assigning the child to tutoring or not, and the child's causal effect is $Y_i(1) - Y_i(0) = 1 - 0 = 1$ regardless of treatment assignment. Randomnesss in our data comes from treatment assignment and, from a population perspective, sampling.

2. **Potential outcomes are implicitly defined in some context, with everything else (pre-treatment) held constant.** Each run of an experiment in some place and time has its own associated potential outcomes, which may reflect various circumstantial factors. In the tutoring experiment, it is **never** possible to observe $Y_i(0)$ and $Y_i(1)$ because I would have to observe the same child twice, with every single pre-treatment variable in the universe the same *except* tutoring vs not tutoring. For example, if I give a child a pre-test in September and then give the same test in June after a year of tutoring, then $Y_{i,Sept}$ is not $Y_{i,June}(0)$.

3. **Potential outcomes have a distribution across individuals** This distribution comes from variation in a sample or population and $Y(t)$ may be treated as random variables from that perspective. E.g., if the experiment has 100 children and 80 of them have $Y(1) = 1$ while 20 have $Y(1) = 0$, then the mean potential outcome under treatment is $.80(1) + .20(0) = .80$ and we could also consider (in theory) the joint distribution of $Y_i(0), Y_i(1)$ and whether treatment assignment is independent of the potential outcomes: $\{Y_i(0), Y_i(1)\}_{i=1}^n \perp T_i$

4. **Causal effects require counterfactuals.** To say treatment T caused Y, it must be possible to imagine a unit being assigned an alternative treatment (e.g., no tutoring). This is why causal effects of characteristics such as race, sex, and age are sometimes debated. Does it make sense to speak of $Y_i(r)$ (the outcome individual i would have had if assigned race r) given that no experimenter can set a person to a different race in the same way as assigning tutoring or not? Some experiments are really looking at perceived characteristics (e.g., audit studies which tweak otherwise identical resumes) but that does not resolve the issue in all cases.[1]

5. **Causes of effects vs effects of causes**: our focus is on 'effects of causes' (will eating a cookie make me happy?) rather than searching for the 'causes of effects' as a detective does (who stole my cookie?). 'Cause of effects' questions are not the focus of this course but can be considered in a potential outcomes framework (see village autonomy example in Module 1 lecture slides). Importantly, even a strong correlation between a condition and an outcome (usually when we observe villages that revolt, they are autonomous) does not guarantee that the condition is necessary or sufficient for the outcome to occur – there may be other factors at play that are not balanced between the 'treated' (has the condition) and untreated (does not) group. Studying 'effects of causes' can be easier because we start with an treatment that is assigned in some random fashion so that differences between treated and untreated groups balance out on average. As we will see, this can justify arguing that observed associations are causal.

1.5 Further Resources

1. Blog post by Guillaume Basse and Iav Bojinov: Introduction to the Potential Outcomes Framework https://www.causalconversations.com/post/po-introduction/

2. Matt Masten Causal Inference Bootcamp (brief non-technical overview videos of some topics we'll cover this semester; video 2.19 gives motivation for module 1 pset) https://mattmasten.github.io/bootcamp/

[1] See, for example, https://www.phenomenalworld.org/analysis/direct-effects/ for more discussion of these kinds of issues.

Modules 2-5: Causal Identification given Randomized Treatments (or at least Randomized Encouragements)

Running Example: Tutoring Experiment: Suppose we run an experiment at a school to see if a tutoring program causes children to pass a math test. We have binary treatment $T = 1$ (tutoring) and $T = 0$ (no tutoring) and binary outcomes $Y = 1$ (pass math test) and $Y = 0$ (does not pass math test). If it makes more sense to think of a continuous outcome, we'll let Y be the test score. There are n total children in the experiment with n_1 treated and n_0 not treated.

2 Module 2: Sharp Null and Permutation Tests

Big Picture: We do within-sample inference for hypotheses about individual-level effects in an experimental setting. In particular, we consider sharp null hypotheses, which are hypotheses such that when true, all potential outcomes are known. We use a permutation test approach to decide whether it is plausible our observed data were generated under that null.

2.1 Key Concepts

1. **Assignment Mechanism** determines which units receive which treatment. Knowledge of this is crucial for being able to identify and estimate causal effects.

2. **Unconfounded Assignment**: when treatment assignment does not depend on the set of potential outcomes: $\{Y_i(1), Y_i(0)\}_{i=1}^n \perp T$. E.g., if all children with $Y_i(0), Y_i(1) = (1,1)$ are assigned tutoring and all with $Y_i(0), Y_i(1) = (0,0)$ are assigned no tutoring, then assignment is confounded. It will look as if tutoring makes a difference when each unit-level effect is 0.

3. **Completely Randomized Design (CRD)**: an unconfounded assignment mechanism that randomly selects n_1 to receive treatment and n_0 to receive control. Each selection of n_1 treated units has probability $\binom{n}{n_1}^{-1}$ and $P(T_i = 1) = \frac{n_1}{n}$ for all units. Note that assignments are not independent because we require n_1 treated and n_0 control. We have unconfounded assignment because all configurations of n_1 treated and n_0 control are equally likely, but $T_i \not\perp\!\!\!\perp T_j$ since $P(T_i = 1|T_j = 1) = \frac{n_1-1}{n-1} \neq \frac{n_1}{n} = P(T_i = 1)$.

4. **Bernoulli Randomization**: an unconfounded assignment mechanism where treatments are assigned by independently sampling from a Bernoulli(p) (i.e., flipping a coin with probability p of heads for each unit). This is **not** the same as CRD. In CRD, assignments are not independent because we require n_1 treated and n_0 control. In Bernoulli randomization, $T_i \perp\!\!\!\perp T_j$ and $P(T_i = 1|T_j = 1) = P(T_i = 1) = p$. Here, n_1 and n_0 become random with $n_1 \sim Binomial(n, p)$

5. **Sharp Null/Strong Null**: most generally, this is any null such that under the null, all potential outcomes are known. Commonly, it means the null of no treatment effect for ANY unit $Y_i(1) = Y_i(0)$ for $i = 1, ..., n$. This is stronger than posing that an average effect is 0. In the binary treatment case, it assumes the only principal strata are $(0,0)$ and $(1,1)$. Another common sharp null is the constant additive effect $Y_i(1) - Y_i(0) = \tau \neq 0 \quad \forall i$. Here, we have

$$Y_i(1 - T_i) = Y_i(T_i) + (1 - T_i)\tau - T_i\tau$$

   (Try plugging in $T_i = 1$ and $T_i = 0$ to see why this is true!) However, rejecting such a null may not tell us much. Rejecting $Y_i(0) = Y_i(1)$ tells us only treatment is doing *something for someone* – direct effects, spillover effects, a few units with an effect, many – the test does not say.

6. **Permutation Test/ Fisher Randomization Test** - see below

7. **Fisher Exact Test**: a permutation test in case of binary treatment and outcome. Exact in the sense that can write down reference distribution in closed form.

8. **Wilcox Rank-Sum Tests**: permutation test used for continuous outcomes. Has power against additive shifts in Y between treated and control. S is sum of ranks of the treated Y_i's relative to full pool of Y's. The more Y's for treated units tend to be higher (lower) than for untreated units, the more evidence against null.

2.2 Further Resources

1. Blog post by Guillarume Basse and Iav Bojinov: Randomization-based inference: the Fisherian Approach https://www.causalconversations.com/post/frt/

2. Imbens and Rubin (2015) Chapter 5

2.3 Permutation Test Steps

The Steps:

1. specify sharp null, often of the form $g(Y_i(0), Y_i(1)) = \tau \quad \forall i$ for some function g
2. specify **observable** test statistic $S = f(T, Y, \tau)$, ideally with power against alternatives of interest (though test is valid regardless).
3. randomly permute treatment assignment according to **known** assignment mechanism and calculate $S^\pi = f(T(\pi), Y(T(\pi)), \tau)$ for each permutation π.[2] This gives the statistic we *would have observed* for the data that *would have resulted* for a different assignment in the world where the null is true. For small n, may look at all permutations but in general, sample permutations (Monte Carlo).
4. compare observed statistic S^{obs} to distribution of $S^{(k)}$ calculated from $k = 1, ..., K$ permutations to calculate p-value. If, under the null, the value of the test statistic is expected to be 0, the two-sided p-value can be calculated as

$$p = \frac{1}{K+1}\left(1 + \sum_{k=1}^{K} I(|S^{(k)}| \geq |S^{obs}|)\right)$$

More generally, an approach that works even if the null distribution is not centered at 0 is calculating

$$p_1 = \frac{1}{K+1}(1 + \sum_{k=1}^{K} I(S^{(k)} \geq S_{obs})) \quad p_2 = \frac{1}{K+1}(1 + \sum_{k=1}^{K} I(S^{(k)} \leq S_{obs}))$$

and taking the p-value to be

$$p = 2 * \min(p_1, p_2)$$

The 1+ reflects the fact that we also have our observed test statistic. It prevents the p-value from ever being 0 (extra: $p = 0$ would violate a standard theoretical requirement for p-values called superuniformity).

Example 1: suppose complete randomization for tutoring experiment. Given null of no effect for any child, we might use

$$S_{obs} = f(T, Y) = \frac{1}{n_1}\sum_{i=1}^{n} T_i Y_i - \frac{1}{n_0}\sum_{i=1}^{n}(1 - T_i)Y_i$$

which will be 0 on average under null since the principal strata $(1, 1)$ and $(0, 0)$ tend to be balanced between treat and control. We would then permute treatment by drawing a permutation π and a new treatment vector $T_{\pi(1)}, ..., T_{\pi(n)}$. Since under this null $Y(T_i) = Y(T_{\pi(i)})$, we can calculate

$$S_\pi = \frac{1}{n_1}\sum_{i=1}^{n} T_{\pi(i)} Y_i - \frac{1}{n_0}\sum_{i=1}^{n}(1 - T_{\pi i})Y_i$$

If we permute treatment, this breaks any relationship of T to Y. If there was no relationship between T and Y to begin with, S^{obs} will not be unusual relative to permutation versions. S^{obs} being unusual is evidence of there being unit(s) with an effect of tutoring (from the $(0, 1)$ or $(1, 0)$ strata)

Example 2: what if now instead, we pose as our null that $Y_i(1) - Y_i(0) = \tau$ for all children? We might then use

$$S_{obs} = f(T, Y, \tau) = \frac{1}{n_1}\sum_{i=1}^{n} T_i Y_i - \frac{1}{n_0}\sum_{i=1}^{n}(1 - T_i)(Y_i + \tau) \quad = \quad \frac{1}{n_1}\sum_{i=1}^{n} T_i Y_i(1) - \frac{1}{n_0}\sum_{i=1}^{n}(1 - T_i)(Y_i(0) + \tau)$$

[2]Notation: it is standard to denote permutations by π. π is a bijective function which reorders $\{1, ..., n\}$, sending each index i to a new index. If the permutation moves unit 3 to location 5, we write $\pi(3) = 5$. For example, $Y_3(T_\pi(3)) = Y_3(T_5)$ says that we assign unit 3 to the treatment originally received by unit 5.

since under the null, we expect $Y_i(1) = Y_i(0) + \tau$, we expect this to be 0 on average under the null.**Careful!** You might be tempted to calculate S_π by just permuting the treatments and re-calculating. However this time, we need to be careful about the Y_i's as well. We need to calculate $Y_i(T_{\pi(i)})$ and this takes a bit more careful thinking! For example, if $T_i = 1$ and $T_{\pi(i)} = 0$, then we observe $Y_i = Y_i(1)$. We need to calculate $Y_i(T_{\pi(i)}) = Y_i(0)$, which we would not know in general. However, under the sharp null, we know that if $T_i = 1$, $Y_i(0) = Y_i(1) - \tau = Y_i - \tau$. Can you fill in the other cases in the table below?

T_i	$T_{\pi(i)}$	$Y_i(T_{\pi(i)})$
1	0	$Y_i - \tau$
1	1	?
0	1	?
0	0	?

Using the results of this table, the test statistic under permutation is:

$$S^\pi = \frac{1}{n_1}\sum_{i=1}^{n} T_{\pi(i)} Y_i(T_{\pi(i)}) - \frac{1}{n_0}\sum_{i=1}^{n}(1 - T_{\pi(i)})(Y_i(T_{\pi(i)}) + \tau)$$

Tips on Permutation Tests

1. **Original randomization of treatment vs random permutations of treatment**. The permutations drawn when running the permutation test must mimic the original assignment mechanism that generated the data. Given complete randomization, permutations are drawn at random over all $n!$ permutations of $(T_1, ..., T_n)$. Given stratified randomization, we should only permute treatments within strata.

2. **Permutation test can be valid, even in presence of spill-over effects, clustering etc.**: under the sharp null, all potential outcomes are known. Your null could incorporate assumptions about spillover, no spillover etc. - the test will still be valid (though what you reject if you reject the null will depend on what your null is...). The **key** thing is that you know how randomization was done and permute according to that.

2.4 Clarification on the Rank-Sum Test; Inverting Tests

Rank-Sum Test in General (outside of causal inference)

The rank-sum test is used outside of the causal inference setting. The general set-up supposes that we have observations $X_1, ..., X_m$ i.i.d. from distribution F and, independently, $Y_1, ..., Y_n$ i.i.d. from distribution G. Suppose we assume that G and F are the same exact distribution except for possibly a mean shift τ such that $Y_j - \tau \sim F$ for some value τ. Given that assumption, we can test hypotheses about the value of τ. Suppose we hypothesize $H_0 : \tau = \tau_0$. Then we calculate the rank-sum test statistic as follows

1. Let $(Z_1, ..., Z_{n+m}) = (X_1, ..., X_m, Y_1 - \tau_0, ..., Y_n - \tau_0)$
2. Let R_i be the rank of Z_i (rank relative to the entire set of pooled observations)
3. Calculate the test statistic

$$S_{\tau_0} = \sum_{i=n+1}^{n+m} \mathrm{R}_i(Z_i)$$

Under the null distribution, the Z_i's should be i.i.d. (Y's indistinguishable from X's) and the distribution of S_{τ_0} will be exactly as derived in Slide 11.

How does this relate to the Rank-Sum Test used in causal inference?

The general rank-sum test above is framed from a super-population perspective. We make no sharp null hypothesis about individual observations – only about a mean shift. In principle, we could also do such a test in a causal setting for a null hypothesis of the form $H_0 : \mathbb{E}(Y(1)) = \mathbb{E}(Y(0)) + \tau_0$ describing mean shift between the treated and control potential outcome. That is, we could:

1. Let F be the population distribution of potential outcomes under control
2. Let G be the population distribution of potential outcomes under treatment
3. Assume the population distributions of potential outcomes under treatment and control differ only by a mean shift τ
4. Assume we have a random sample from the population and complete randomization so that the observed control units $\{Y_i : T_i = 0\}$ are an i.i.d. sample from G and the observed treated units $\{Y_i : T_i = 1\}$ are an i.i.d. sample from F
5. Hypothesize mean shift τ_0 and calculate S_{τ_0} as above by calculating the ranks of $Z_i = Y_i - \tau_0 T_i$ for all observations and then summing only the ranks of the treated group.

If the assumptions hold, this is a valid test for a difference in mean outcomes under treatment and control

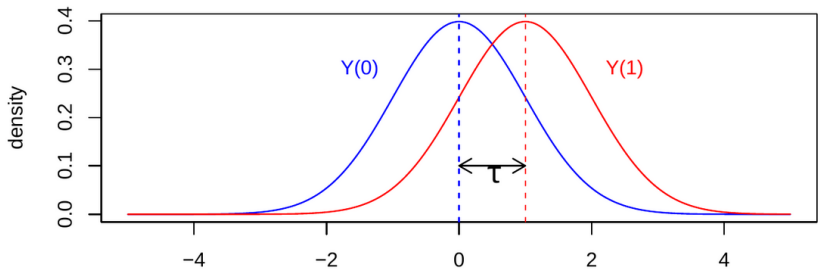


Figure 1: Illustration of mean shift

How does this relate to using the rank-sum statistic in a permutation test?

The permutation test use of the rank-sum test is different from the super-population approach above.

1. Instead of a super-population perspective (thinking about population distributions), it focuses only on the particular sample of units we have observed. It does not assume anything about population distributions or sampling. The test is valid as a permutation test even if we have an extremely unrepresentative sample of some larger population and even if the treatment and control group in the population have wildly different distributions (not just a mean shift)

2. We generally use it to test a **sharp null** of constant additive effect $Y_i(1) - Y_i(0) = \tau$ at the mean level.[3] This does also imply that under the null, the sample means are shifted by τ but the reverse is not true.

3. The rank-sum test statistic is not special in that we could have chosen other test statistics to test whether our hypothesized additive shift holds. Those other statistics might have more or less power to detect violations of additive shift

That said, if we pick the rank sum statistic for our permutation test, the test statistic calculated and the distribution we compare it to is **exactly the same** as if we run the difference in means test.

So what is the difference really?

It comes down to what assumptions you make. The two tests are algebraically the same but make different sets of assumptions. If you are willing to assume that you have an i.i.d. sample from a super-population, complete randomization, and that the treatment and control distributions differ only by mean shift, then rejecting a null of $\tau = \tau_0$ indicates the mean shift is not τ_0 (or one of your assumptions is wrong...). If you want to assume nothing about how your data were sampled but you are willing to assume complete randomization and a constant additive shift, then rejecting indicates your hypothesized additive shift τ_0 is false (in the case of $Y_i(0) = Y_i(1)$ hypothesis, it simply reflects at least some units having a treatment effect of some magnitude). In any application, you will need to argue which assumptions are plausible.

General Comment: A hypothesis test always poses both (1) assumptions and (2) the null and alternate hypotheses. If (1) hold, then a small p-value is evidence against the null. Otherwise, a small p-value could reflect a false null, a false assumption, or both. In the permutation test, the treatment assignment mechanism (in the simplest case, complete randomization) is an assumption while the sharp null is the hypothesis. If we assume complete randomization but this is false, it could be that a sharp null DOES hold but we observe an extreme S^{obs} because of confounding.

2.4.1 Inverting the Permutation Test

Permutation tests do not tell directly us about effect size, but there are ways to use them to get a point estimate and confidence interval (set). Broadly:

- Form a **confidence set** by collecting all the null hypothesized values that would not be rejected by the test. If our test statistic monotonic in τ, this will be a confidence interval. For example, for the rank sum test, we would calculate the test statistic and p-value for a range of $H_0 : \tau = \tau_0$ and find the set of τ such that we fail to reject.

- Obtain a **point estimate** by finding the value that yields the largest p-value (least evidence against the null). For example, for the rank-sum test, because the null distribution is symmetric, this corresponds to finding the hypothesized τ_0 such that the test statistic calculated under that shift is equal to the mean under the null distribution ($\frac{n_1(n+1)}{2}$). If we cannot get exact equality, we average the τ that get us close from above and below.

The discussion above again applies here. We can invert the rank-sum test either way, but whether our confidence interval and point estimates represent the value of a **population difference in means** or of a **constant individual-level additive effect in the finite sample** depends on what assumptions we make.

[3] Other sharp nulls could also be tested, but the test statistic might not have a lot of power to reject some of them.

2.5 Extension: Conditional Randomization Tests

2.5.1 Motivation: Spillover Effects

Conditional Randomization Tests (CRT) are form of permutation test especially relevant for detecting **spillover effects** in the context of network data. Recall that spillover occurs when the treatments of other units affect unit i's outcome. For example, for two units, if unit j's treatment has a spillover effect on unit i, unit i's potential outcome becomes $Y_i(T_i = t, T_j = t')$. For binary treatment, the spillover effects of unit j on i are then:

$$Y_i(T_i = t, T_j = 1) - Y_i(T_i = t, T_j = 0) \qquad \text{for } t = 0, 1$$

while the direct effects are

$$Y_i(T_i = 1, T_j = t) - Y_i(T_i = 0, T_j = t) \qquad \text{for } t = 0, 1$$

this can of course be generalized beyond pairs of units and binary treatments. In some contexts, we might loosen the no-spillover effect assumption without getting rid of it entirely by, for example, assuming spillover comes only from people directly connected to unit i in a network or only within certain blocks (families, neighborhoods etc.). The challenge of spillover effects in general is that it greatly expands the number of possible treatments because each possible configuration of individual treatments counts as a different treatment.

2.5.2 Example 1 (Toy Example from Lecture)

In a CRT, we condition on some treatment assignments while permuting others to test independence between the permuted treatments and the outcomes of some set of focal units, *conditional* on (controlling for) the treatments of the non-permuted units. Consider the following network from lecture:

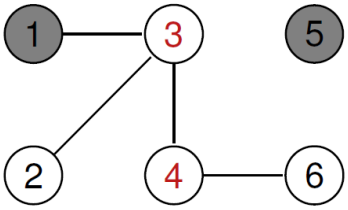


Figure 2: Network from lecture slides. Grey units are treated in the original experiment. White units are not.

We might ask, "Are units 3 and 4 affected by the treatment of their friends?" **Assume** randomized assignment of treatment so that correlations between outcome and treatment reflect causal relationships. Our null hypothesis is that there are no spillover effects and hence

$$Y_3, Y_4 \perp T_1, T_2, T_4, T_6 | T_3, T_4$$

Under this null, there could still be direct effects of treatment such that $Y_3, Y_4 \not\perp T_3, T_4$ or even spillover effects within the focal group (T_3 on Y_4 or vice versa). The null only says that *conditional* on any direct treatment effects, there is no further spillover effect. To test the hypothesis, we measure the correlation between the outcomes of 3 and 4 and the treatment of their friends and see whether the observed correlation is unusual relative to the correlation we observe when we randomly permute treatment across their friends and non-friends.

Of course, this is a tiny toy example, but you can imagine having a much larger dataset on friendship networks and a test statistic reflecting the correlation between unit i's outcome and unit i's friends' treatments. Imagine, for example, that Y_i is a healthy behavior that experimenters would like to encourage via treatment. If treated units with treated friends tend to adopt the behavior much more often than treated units with untreated friends (and similarly for untreated units with treated/untreated friends), then when we permute the friends' treatments we should 'break' that correlation. This means that in our permuted data, we'll have some mix of all combinations of behavior-adopting and non-adopting units with few or many friends treated and there won't be a correlation any more! Think of this as getting a baseline. We might observe some correlation between number of friends treated and healthy behavior adoption, but is that correlation large enough to conclude it's unlikely to be just by chance? How large is large? The permutation test helps us gain some point of comparison.

2.5.3 Formal Set-up

The CRT is used beyond causal inference to test hypotheses of the form

$$H_0 : Y \perp\!\!\!\perp X|Z$$

The key requirement is that $p(X|Z)$ is known. We define some test statistic $S = f(X, Y, Z)$ that is sensitive to violations of the independence assumption. The test then proceeds by

1. Sampling m draws of $\tilde{X}$ from $p(X|Z)$ ($\tilde{X}$ is a vector of same length as X in the original data)
2. Computing test statistic for each draw $\tilde{S}^{(j)} = f(\tilde{X}^{(j)}, Y, Z)$
3. Computing p-value to reflect how unusual S_{obs} is compared to the $S^{(j)}$ as in the standard permutation test

2.5.4 Set-up applied to testing spillover in a completely randomized experiment

Given an experiment with units $1, ..., n$, we define the following two partitions of the units

1. **Focal units** F - outcomes are of interest; **Non-Focal Units** F^c - outcomes not of interest
2. **Permute-treatment Units** R; **Fix-treatment Units** R^c

Although when testing for spillover, the focal units will be in the fix-treatment group, the set of focal units and set of fix-treatment units are not always exactly the same. In the Module 2 problem set, you will see an example where the focal units are somewhat trivially part of the fix-treatment group because they are not eligible for treatment, but we also fix the treatments of a further group of units. The null hypothesis of the form $Y \perp\!\!\!\perp X|Z$ is now:

$$H_0 : \{Y_i : i \in F\} \perp\!\!\!\perp \{T_i : i \in R\}|\{T_i : i \in R^c\}$$

If this hypothesis is true, then it would indicate no spillover effects from the treatment of units in set R on the outcomes of the units in F. This does not imply that there are no spillover effects at all. There could, for example, be spillover effects within fixed-treatment group. Our procedure is then to

1. Choose test statistic sensitive to relationships of interest between the outcomes of F and treatments of R
2. Permute the treatment labels m times within the R group only
3. For each of the m permutations, re-calculate test statistic $S^{(j)}$ under permutation π_j
4. Evaluate how unusual $S_{observed}$ is compared to the $S^{(j)}$'s – i.e. calculate p-value as above

Clarification 1: in a completely randomized design, permuting the treatment labels in R while fixing those in R^C (Step 2) is equivalent to drawing from $p(X|Z)$ as in the general set-up where $X = \{T_i : i \in R\}$ and $Z = \{T_i : i \in R^C\}$. This is because in completely randomized design, we draw from one of $\binom{n}{n_1}$ possible assignments uniformly at random. If we fix some k assignments including $s \leq k$ treatment $T_i = 1$ assignments, the conditional distribution of the rest of the assignments is a uniform random draw over $\binom{n-k}{n_1-s}$ assignments to the remaining units.

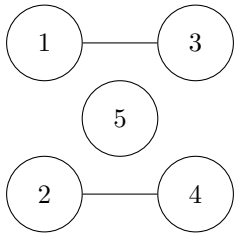
Clarification 2: what does this look like in terms of potential outcomes? Let t_S represent a vector of treatments for some set S and let T_S be the observed treatments for S. In finite sample terms, the null hypothesis above is equivalent to;

$$H_0 : Y_i(t_R, t_{R^c} = T_{R^c}) = Y_i(t'_R, t_{R^c} = T_{R^c}) \quad \text{for all } t'_R, \text{ and all } i \in F$$

Note that it is also be possible, though a bit harder to think about, to pose a *distributional* hypothesis about $P(Y_i(t_R, t_{R^c} = T_{R^c}))$ and $P(Y_i(t'_R, t_{R^c} = T_{R^c}))$ for some defined population. Specifying this population and how we sample from it is important here since individuals have networked relationships and the presence or absence of one individual in a sample may change another individual's potential outcomes.

2.5.5 Example 2 (Some Difficulties of Interpretation)

Suppose we have the following friendship network among 5 people:



We might be interested in whether 3 has a spillover effect on 1, suggesting the hypothesis:

$$Y_1 \perp\!\!\!\perp T_3 | T_1, T_2, T_4, T_5$$

and whether 4 has a spillover effect on 2, suggesting

$$Y_2 \perp\!\!\!\perp T_4 | T_1, T_2, T_3, T_5$$

If we try to test these at the same time, we have different fixed-treatment sets! If we pool the conditioning sets, we end up with T_1, T_2, T_3, T_4, T_5 and no treatments left to permute. One solution would be to instead condition on T_1, T_2, T_5 and permute T_3, T_4. However, **this impacts the interpretation**. If we reject the null

$$Y_1, Y_2 \perp\!\!\!\perp T_3, T_4 | T_1, T_2, T_5$$

Then we have not demonstrated that there is a spillover effect from 3 to 1 and from 4 to 2. There could also or instead be an effect of 4 on 1 or 3 on 2. Hence we can only conclude there is some spillover effect of the group $\{3, 4\}$ on the group $\{1, 2\}$. This is again a toy example but it represents the bigger idea that if we want to test for spillover effects on multiple individuals (necessary in practice to be able to detect some correlation) we often can't construct the test in such a way that rejecting it conclusively demonstrates that spillover effects are coming only from people's direct friends and not from friends of friends or just other people in the network. We might have a test statistic which is sensitive to there being some relationship to friends' treatment, and we could argue based on the context that it is reasonable to assume that if there is spillover, it is **only** from direct friends, but just rejecting the null of our test does not tell us that "only."

2.5.6 Example 3 (Choice of focal group affects nature of conclusions)

Imagine we have an experiment which includes parents and their children (suppose for simplicity each pair has only one child in the experiment). Suppose we completely randomize treatment.

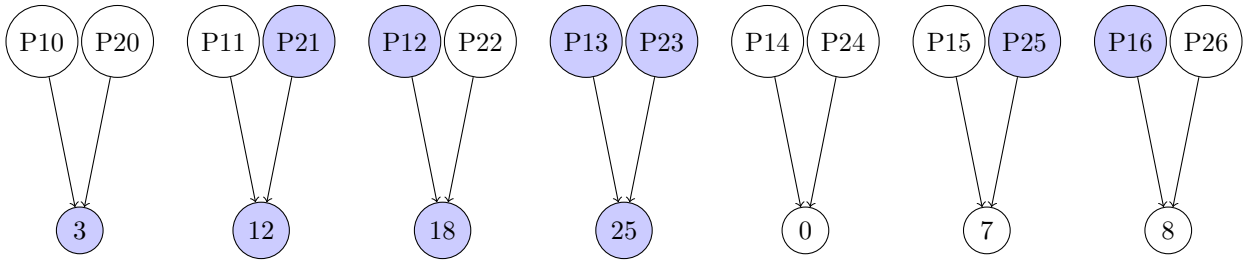


Figure 3: Parent-child network: The upper row is parents. The lower row is children. Blue represents treated units and white represents control units. The numbers in the child nodes represent their outcomes.

Suppose we want to study the spillover effects of parent treatment on child. Then we should make child our focal units, fix child treatments, choose a test statistic sensitive to how child's outcome relates to parent treatment, and randomly permute parent treatments (we might also do this just within treated children or just within untreated children). In the diagram above, I have set it up so that children with more treated parents tend to have higher outcomes. If we permute parent labels, then this will break that connection. Suppose instead we want to study spillover effects of child treatment on parent outcomes. Now we should treat parents as our focal group and permute treatments of children! In this way, the choice of focal group can have a big impact on what our test represents.

3 Module 3: Average Treatment Effects

Big picture: We move to within-sample and population-level inference for average effects (Neyman approach), rather than testing hypotheses about presence of individual causal effects (Fisher's approach). We are still in an experimental setting. Averaging means that individual-level effects may be heterogeneous or even cancel out (average effect being 0 need not imply individual ones are) and population-level inference brings in more **asymptotic** and **sampling** considerations.

3.1 Within Sample vs Population Perspective

1. **Within-sample perspective**: given data on n units, each individual unit has a causal effect $Y_i(1) - Y_i(0)$ and we might be interested in these or in the average effect $\frac{1}{n}\sum_{i=1}^{N}(Y_i(1) - Y_i(0))$ *on those units*. The only source of **randomness** within a sample is the **treatment assignment** determining which outcomes are revealed. In Module 2, Fisher Permutation tests were *only* testing a hypothesis about within-sample effects for a fixed group of people.

2. **Population Perspective**: Arguing that an estimate of a sample-level average causal quantity is also an estimate of the average effect for an entire (possibly infinite) population requires arguing the sample generalizes to that population, e.g., because it is a simple random sample. From this perspective, there is randomness both from treatment assignment and sampling.

	Within-sample inference	**Population inference**
Justifications	Based on fixed, finite sample only	Finite sample, super-population sampling, and sometimes asymptotic (as $n \to \infty$) arguments
Sources of randomness	Treatment randomization	Treatment randomization, Sampling
Applicable methods (so far)	Fisher Permutation tests, Neyman SATE	Neyman PATE
Tutoring Example	100 children in a given school and only want to study if tutoring affects their math scores. If $\bar{Y}_1 = .70$, that is the only relevant treatment mean outcome.	random sample of 100 children from a given school and we want to study if tutoring affects math scores of all students in the school. Sample might have $\bar{Y}_1 = .70$ while (unknown to us) population mean is $\mathbb{E}(Y(1)) = .80$.

Notation: we distinguish the sample vs population perspective by conditioning on $\mathcal{O}_n = \{Y_i(0), Y_i(1)\}_{i=1}^n$ in the sample case. Conditional on $\mathcal{O}_n$, for example, we have $\mathbb{E}(Y_i(1)|\mathcal{O}_n) = Y_i(1)$ because $Y_i(1)$ is a fixed (possibly unobserved) potential outcome for a unit in our sample. Without conditioning on $\mathcal{O}_n$, however, $\mathbb{E}(Y_i(1))$ is the expected value for a population.

3.2 Summary: ATE under Completely Randomized Design

	Sample Average Treatment Effect (SATE)	**Population Average Treatment Effect** (PATE)
Estimand	$\tau = \frac{1}{n}\sum_{i=1}^{n}(Y_i(1) - Y_i(0)) = \overline{Y(1)} - \overline{Y(0)}$	PATE $= \mathbb{E}(Y_i(1) - Y_i(0))$
Identification strategy	(1) randomized assignment of treatment so $T \perp\!\!\!\perp Y(0), Y(1)$, which implies identification equality $\mathbb{E}(\hat{\tau}\|\mathcal{O}_n) = \tau$	(1) and (2) having a simple random sample (i.i.d. draws) so that $\mathbb{E}(\tau) = \mathbb{E}(Y_i(1) - \mathbb{E}(Y_i(0))$
Estimator	**Difference in Means Estimator** $\hat{\tau} = \frac{1}{n_1}\sum_{i=1}^{n} T_i Y_i - \frac{1}{n_0}\sum_{i=1}^{n}(1-T_i)Y_i = \bar{Y}_1 - \bar{Y}_0$	
Variance of Estimator	$\mathbb{V}(\hat{\tau}\|\mathcal{O}_n) = \frac{1}{n}\left(\frac{n_0}{n_1}S_1^2 + \frac{n_1}{n_0}S_0^2 + 2S_{01}\right)$ Theoretical bounds from Cauchy-Schwartz $\frac{n_0 n_1}{n}\left(\frac{S_1}{n_1} \pm \frac{S_0}{n_0}\right)^2$	$\mathbb{V}(\hat{\tau}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n_0}$
Estimate of Variance	Unidentifiable because S_{01} unidentifiable. Can estimate bound or use PATE variance estimator since it is conservative on avg.: $\mathbb{V}(\hat{\tau}\|\mathcal{O}_n) \leq \mathbb{E}(\hat{\mathbb{V}}(\hat{\tau})\|\mathcal{O}_n)$ with equality iff $Y_i(1) - Y_i(0) = c \ \ \forall i$	$\hat{\mathbb{V}}(\hat{\tau}) = \frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_0^2}{n_0}$ Unbiased estimator of $\mathbb{V}(\hat{\tau})$
Asymptotics for obtaining confidence intervals and p-values (Slide 12)	Finite sample CLT applies but often use conservative variance PATE version in practice	Consistent, asymptotically normal as $n \to \infty$ with $\frac{n_1}{n}$ constant. $\frac{\hat{\tau} - \tau}{\sqrt{\mathbb{V}(\hat{\tau})}} \xrightarrow{d} N(0,1)$

$$S_t^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i(t) - \overline{Y(t)})^2 \qquad S_{01} = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i(0) - \overline{Y(0)})(Y_i(1) - \overline{Y(1)})$$

$$\sigma_t^2 = Var(Y_i(t)) = \mathbb{E}(S_t^2) \qquad \hat{\sigma}_t^2 = \frac{1}{n_t - 1}\sum_{i=1}^{n} I(T_i = t)(Y_i - \bar{Y}_t)^2$$

$$\overline{Y(t)} = \frac{1}{n}\sum_{i=1}^{n} Y_i(t) \qquad \bar{Y}_t = \frac{1}{n_t}\sum_{i=1}^{n} I(T_i = t)Y_i$$

3.2.1 Detailed logic for identification of SATE and PATE (see also Slide 7)

First, because whenever $T_i = 0$, $T_iY_i = T_iY_i(1) = 0$ and whenever $T_i = 1$, $Y_i = Y_i(1)$ (and vice versa with $Y_i(0)$),[4]

$$\hat{\tau} = \frac{1}{n_1}\sum_{i=1}^{n} T_iY_i - \frac{1}{n_0}\sum_{i=1}^{n}(1-T_i)Y_i = \frac{1}{n_1}\sum_{i=1}^{n} T_iY_i(1) - \frac{1}{n_0}\sum_{i=1}^{n}(1-T_i)Y_i(0)$$

Next, conditioning on the particular set of potential outcomes in our sample $\mathcal{O}_n$, we have

$$\begin{aligned}
\mathbb{E}(T_iY_i(1)|\mathcal{O}_n) &= Y_i(1)\mathbb{E}(T_i|\mathcal{O}_n) && \text{(because conditioning on } \mathcal{O}_n\text{)} \\
&= Y_i(1)P(T_i = 1|\mathcal{O}_n) && \text{(expectation of an indicator)} \\
&= Y_i(1)P(T_i = 1) && \text{(unconfounded assignment mechanism)} \\
&= Y_i(1)\frac{n_1}{n} && \text{(completely randomized design)}
\end{aligned}$$

Similarly, $\mathbb{E}((1-T_i)Y_i(0)) = Y_i(0)\frac{n_0}{n}$ and plugging this in yields

$$\mathbb{E}(\hat{\tau}|\mathcal{O}_n) = \frac{1}{n_1}\left(\frac{n_1}{n}\sum_{i=1}^{n} Y_i(1)\right) - \frac{1}{n_0}\left(\frac{n_0}{n}\sum_{i=1}^{n} Y_i(1)\right) = \frac{1}{n}\sum_{i=1}^{n}(Y_i(1) - Y_i(0))$$

Finally, we use the law of total expectation and the assumption that we have an i.i.d. sample to reason about the PATE

$$\mathbb{E}(\hat{\tau}) = \mathbb{E}(\mathbb{E}(\hat{\tau}|\mathcal{O}_n)) = \mathbb{E}(\tau) = \mathbb{E}(\frac{1}{n}\sum_{i=1}^{n} Y_i(1) - Y_i(0)) = \mathbb{E}(Y_i(1)) - \mathbb{E}(Y_i(0)) = PATE$$

3.2.2 Alternative form of $\mathbb{V}(\hat{\tau}|\mathcal{O}_n)$: why $\hat{\mathbb{V}}(\hat{\tau})$ is conservative

Above we have $S_{01} = \text{Cov}(Y_i(0), Y_i(1)|\mathcal{O}_n)$. Now define the sample variance of the unit level effects $Y_i(1) - Y_i(0)$ as:

$$S^*_{01} := V(Y_i(1) - Y_i(0)|\mathcal{O}_n) = \frac{1}{n}\sum_{i=1}^{n}((Y_i(1) - Y_i(0)) - \tau)^2$$

Then using the relation

$$S^*_{01} = \mathbb{V}(Y_i(1)|\mathcal{O}_n) + \mathbb{V}(Y_i(0)|\mathcal{O}_n) - 2\text{Cov}(Y_i(0), Y_i(1)|\mathcal{O}_n) = S_1^2 + S_0^2 - 2S_{01}$$

we have[5]

$$\mathbb{V}(\hat{\tau}|\mathcal{O}_n) = \frac{1}{n}\left(\frac{n_0}{n_1}S_1^2 + \frac{n_1}{n_0}S_0^2 + 2S_{01}\right) = \frac{1}{n}\left(\frac{n_0}{n_1}S_1^2 + \frac{n_1}{n_0}S_0^2 + S_1^2 + S_0^2 - S^*_{01}\right) = \frac{S_1^2}{n_1} + \frac{S_0^2}{n_0} - \frac{S^*_{01}}{n}$$

This form more clearly illustrates why the PATE estimator is in expectation conservative since $S^*_{01} \geq 0$ and

$$\mathbb{E}(\hat{\mathbb{V}}(\hat{\tau})|\mathcal{O}_n) = \mathbb{E}(\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_0^2}{n_0}|\mathcal{O}_n) = \frac{S_1^2}{n_1} + \frac{S_0^2}{n_0} \geq \mathbb{V}(\hat{\tau}|\mathcal{O}_n)$$

It also shows that this inequality is equality if and only if $S^*_{01} = 0$, which occurs only when there is a constant additive treatment effect as in the sharp null.

Note: See also question 5 of the Module 3 Review Sheet for connections between the two theoretical variances $\mathbb{V}(\hat{\tau})$ vs $\mathbb{V}(\hat{\tau}|\mathcal{O}_n)$ derived via the Law of Total Variance.

[4] More formally: $T_iY_i = T_i\ (Y_i(1)T_i + Y_i(0)(1-T_i)) = T_i^2Y_i(1) + T_i(1-T_i)Y_i(0) = T_iY_i(1)$. This calculation is also done here.
[5] Source: Imbens and Rubin pg. 89. See also Module 3 review sheet for derivation of the original form.

3.3 More on identification

> *"Statistical inference teaches us "how" to learn from data, whereas identification analysis explains "what" we can learn from it... Koopmans (1949) coined the term "identifiability" and emphasized a 'clear separation between problems of statistical inference arising from the variability of finite samples, and problems of identification in which [the statistician] explore[s] the limits to which inference even from an infinite number of observations is subject.'"* (Basse and Bojinov, 2020, Introduction)

Identification has to do whether we have the right kind of observed information to learn about a quantity of interest (an estimand). In causal inference, it is easy to find estimands that cannot be identified. Most simply, $Y_i(1) - Y_i(0)$, the individual-level treatment effect, is non-identifiable. It is tricky to formalize what identification means in all contexts (optional: see Basse and Bojinov (2020) for an effort to do so). However an intuition is that identification is like requiring there to be a one-to-one relationship between the values an estimand can take and the information contained by observed data in the limit so that as we learn more data, only one value of the estimand will ultimately be compatible with the data. Think of a detective gathering evidence until finally, only one suspect is left!

Parametric identifiability: a statistical model $\{p_\theta(x) : \theta \in \Theta\}$ is identifiable if for any θ_1, θ_2 such that $p_{\theta_1}(x) = p_{\theta_2}(x)$ for all x, we have $\theta_1 = \theta_2$. Intuitively if it were possible to know $p_\theta(x)$ for every single x and *still* multiple θ_i would be compatible, then we cannot ever identify which θ is the true one. For example, suppose $X_1, ..., X_n \overset{iid}{\sim} N(\alpha + \beta, 1)$. Then $\alpha = 2, \beta = 2$ and $\alpha = 3, \beta = 1$ result in the exact same probability distributions! In terms of data, consider the fact that the Law of Large Numbers, $\bar{X}_n \xrightarrow{p} \alpha + \beta$ as $n \to \infty$. Suppose we collect infinite data and learn that $\alpha + \beta = 4$. Since both $\alpha = 2, \beta = 2$ and $\alpha = 1, \beta = 3$ (among infinite others) are compatible with this, α by itself is not identifiable.

In this module: the SATE τ and PATE are our estimands but we cannot directly calculate them from observations. However, under completely randomized design, we show that $\mathbb{E}(\hat{\tau}|\mathcal{O}_n) = \tau$. Remember that in the SATE case, the randomness underlying this expectation comes from random treatment assignment. Similar to the 'infinite data' idea above, this expectation tells us that if we could repeatedly run our experiment in parallel universes, for all possible treatment assignments and then average the resulting $\hat{\tau}$'s, we would obtain the true τ. In the PATE case, we add that if we could sample infinitely many observations from the population, we could learn the true population average effect $\mathbb{E}(\tau) = \mathbb{E}(Y_i(1) - Y_i(0))$. Note that identification concerns an infinite data case, so it would be ok to have a biased estimator in finite samples that is still **consistent** for the estimand of interest. The issue of finite sample **estimation error** $|\hat{\tau} - \mathbb{E}(\hat{\tau}|\mathcal{O}_n)|$ for a single finite dataset is distinct from identification.

Non-identifiable: An example of a non-identifiable quantity is the covariance S_{01} between $Y_i(1)$ and $Y_i(0)$. We never observe $Y_i(0)$ and $Y_i(1)$ at once and have no observed information about their joint distribution, so there is no $f(Y_i, T_i)$ such that $\mathbb{E}(f(Y_i, T_i)) \xrightarrow{p} S_{01}$. To see this concretely, imagine $Y_i(0) \overset{iid}{\sim} N(0,1)$ and $Y_i(1) \overset{iid}{\sim} N(0,1)$. Then their joint distribution could take any of the forms in 4. On the left, $S_{01} = 0$. We have many non-zero individual-level treatment effects but they cancel out on average. On the right, we have the sharp null where no one has a treatment effect. The ATE framework described above cannot distinguish these situations! An important point is that **adding more assumptions can make a quantity identifiable**. For example (Slide 10), if we assume a sharp null of constant additive treatment $Y_i(1) - Y_i(0) = c \ \ \forall i$ as in Module 2, then from the previous section, $S_{01}^* = 0$, $S_{01} = \frac{1}{2}(S_0^2 + S_1^2)$ and since S_t^2 is identifiable via $\hat{\sigma}_t^2$, we have identification of S_{01}.

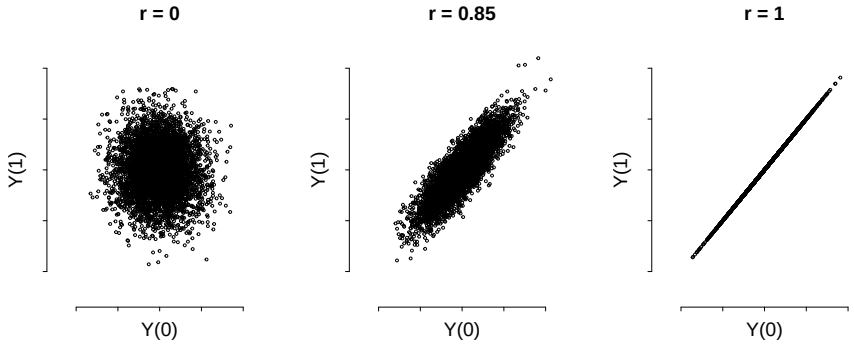


Figure 4: 10,000 samples of $Y(0), Y(1)$ each marginally $N(0,1)$ with varying $r = \text{Corr}(X, Y)$. Generated with `mnormt` package in `R`

Identification is at the heart of causal inference. One way of viewing the modules to come is that they cover a set of different identification strategies for a variety of different estimands.

3.4 Tips

1. **The Law of Total Expectation is your friend**: from a population perspective, we have TWO sources of randomness: one from randomization of treatment, one from sampling. This makes calculating a quantity like $\mathbb{E}(Y_i(1)T_i)$ directly non-straightforward. Again and again in this course, we will leverage the Law of Total Expectation to first deal with one source of randomness (e.g., $\mathbb{E}(Y_i(1)T_i|\mathcal{O}_n)$ to deal only with randomization) and then the next (e.g., $\mathbb{E}(\mathbb{E}(Y_i(1)T_i|\mathcal{O}_n))$ to deal with sampling).

2. **Variance of estimator vs estimate of variance**: in the table above, $\mathbb{V}(\hat{\tau}|\mathcal{O}_n)$ and $\mathbb{V}(\hat{\tau})$ are theoretical quantities. We cannot actually observe $S_0, S_1, S_{01}, \sigma_0^2, \sigma_1^2$. Therefore, these are ALSO estimands and in fact need their own identification strategies and estimators! As noted above, S_{01} is unidentifiable, making the SATE variance unidentifiable in general. Hence the need for a bounds argument.

3. n_0, n_1: are fixed in the experimental design and the results above are valid regardless of their values. However, the choice of n_0, n_1 does affect the variances and hence the precision of our inferences. Ideally, we would pick n_1 and n_0 to reflect the variance of $Y_i(1)$'s and $Y_i(0)$'s. For example, if $Y_i(1)$'s have low variance and $Y_i(0)$'s high variance, we'll get more precision with $n_1 < n_0$ than $n_1 = n_0 = \frac{n}{2}$. Slide 10 gives optimal n_1, n_0 formulas based on minimizing $\mathbb{V}(\hat{\tau}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n_0}$. We cannot know these because they involve σ_1, σ_0 but sometimes researchers do a pilot study to estimate them.

4. **Unconfoundedness and treatment-outcome independence in observed data** are distinct concepts and do not imply each other. Treating $Y_i(t)$'s and T_i's as random variables from a finite or super population perspective,

$$T_i \perp\!\!\!\perp \{Y_i(0), Y_i(1)\}_{i=1}^n \quad \not\leftrightarrow \quad T_i \perp\!\!\!\perp Y_i$$

   but the notation can look similar. $\mathbb{E}(Y_i(1)|T_i = 1)$ and $\mathbb{E}(Y_i|T_i = 1)$ are conceptually distinct, even if they are sometimes equal.

   - $\mathbb{E}(Y(1)|T = 1)$ is a property of the distribution of potential outcomes in some sample and of the assignment mechanism. If the assignment mechanism is unconfounded, then

$$\mathbb{E}(Y_i(1)|T_i = 1) = \mathbb{E}(Y_i(1)|T_i = 0) = \mathbb{E}(Y_i(1)) \qquad \mathbb{E}(Y_i(0)|T_i = 1) = \mathbb{E}(Y_i(0)|T_i = 0) = \mathbb{E}(Y_i(0))$$

   - $\mathbb{E}(Y|T = 1)$ is a property of observed variables. Under unconfoundedness (and consistency - Section 1.2), it is related to the above as follows:

$$\mathbb{E}(Y_i(1)) = \mathbb{E}(Y_i(1)|T_i = 1) = \mathbb{E}(Y_i|T_i = 1) \qquad \mathbb{E}(Y_i(0)) = \mathbb{E}(Y_i(0)|T_i = 0) = \mathbb{E}(Y_i|T_i = 0)$$

     Hence under unconfoundedness, $\mathbb{E}(Y_i(1)|T_i = 1) = \mathbb{E}(Y_i(1)|T_i = 0)$ but whether $\mathbb{E}(Y_i|T_i = 1) = \mathbb{E}(Y_i|T_i = 0)$ is a question of whether $\mathbb{E}(Y_i(1)) = \mathbb{E}(Y_i(0))$ aka **whether there is a treatment effect**. If there tend to be positive unit causal effects, then we may have $\mathbb{E}(Y_i|T_i = 1) > \mathbb{E}(Y_i|T_i = 0)$. E.g., if many students belong to the principal stratum $(0, 1)$ (i.e. only pass test if tutored), then we will have many observed $Y_i = Y_i(1) = 1$ in the treated group and many observed $Y_i = Y_i(0) = 0$ in the control group.

   - Conversely, if the assignment mechanism is confounded, we also may or may not have $\mathbb{E}(Y_i|T_i = 1) = \mathbb{E}(Y_i|T_i = 0)$. For example, if we put all $(0, 1)$ children in the treatment group and all $(1, 0)$ children in the control group. Then $\mathbb{E}(Y_i(1)|T_i = 1) = 1 \neq 0 = \mathbb{E}(Y_i(1)|T_i = 0)$ but $\mathbb{E}(Y_i|T_i = 1) = \mathbb{E}(Y_i|T_i = 0) = 1$.

3.5 Further Resources

1. Blog post by Guillarume Basse and Iav Bojinov: Randomization-based inference: the Neymanian approach https://www.causalconversations.com/post/randomization-based-inference/

2. Blog post by Guillarume Basse and Iav Bojinov: Identification: What is it? https://www.causalconversations.com/post/identification/ – formalizes the notion of identification in an interesting way; this will not come up in the course but might be helpful if you are frustrated by the loose discussion above.

3. Imbens and Rubin (2015) Chapters 6, 9 (Skip 9.6–9.7), and 10 (Skip 10.6–10.7)

3.6 Extension: ATE for Stratified Randomized Designs

The ATE can be identified under various randomization schemes, though the identification equality, estimator, and variance will change. In stratified designs, treatment is assigned randomly within strata. Given n units, J blocks, n_j units in block j with n_{1j} of them treated, with outcomes Y_{ij} for $i = 1, ..., n_j, j = 1, ..., J$ the estimator for the overall average treatment effect is a weighted average of stratum-level Neyman estimators:

Overall Estimand	SATE or PATE as before
Identification Strategy	randomization within strata (and simple random sample if PATE)
Stratum-level effect estimator	$\hat{\tau}_j = \frac{1}{n_{1j}} \sum_{i=1}^{n_j} T_{ij} Y_{ij} - \frac{1}{n_{0j}} \sum_{i=1}^{n_j} (1 - T_{ij}) Y_{ij}$
Overall Estimator	$\hat{\tau} = \sum_{j=1}^{J} w_j \hat{\tau}_j$
Stratum-level Variance Estimator	$\hat{\mathbb{V}}(\hat{\tau}_j) = \frac{\hat{\sigma}^2_{j1}}{n_{1j}} + \frac{\hat{\sigma}^2_{j0}}{n_{0j}}$
Overall Variance estimator	$\sum_{j=1}^{J} w_j^2 \hat{\mathbb{V}}(\hat{\tau}_j)$
Weights	$w_j = \frac{n_j}{n}$

Why do this? stratification can improve the **efficiency** (lower variance) of an experiment. Ideally, strata are groups of units which are similar in their $Y_i(0), Y_i(1)$ values (minimal **within-strata** variance) so that the treated units are good stand-ins for the $Y_i(1)$'s of the control units and vice versa. Of course, we do not know $Y_i(0), Y_i(1)$ and can only define strata using **pre-treatment** covariates which we a priori suspect might be related to outcome similarity. The **across-strata** variance in $Y_i(0), Y_i(1)$ reflects the efficiency gain because we avoid situations where $\hat{\tau}$ takes extreme values not because of large treatment effects but because of chance randomizations that make the treatment and control group contain dissimilar sets of units. See Figure 5 for an illustration.

In fact, it is possible to show that stratification only ever leads to an equal or lower variance - there is never an efficiency loss. Pushed to the extreme, we might even create pairs of similar units with one treated and one control per pair. This is called **matched pairs design** (See Module 3 Gov2003 exercise)

Example: in the tutoring experiment, we might suspect that outcomes (passing math test) are correlated with family income such that wealthier students pass the math test more often, regardless of tutoring. Ideally, the experiment is balanced so that we do not, for example, have only wealthier students assigned to tutoring or vice versa. A completely randomized design, while valid, will only give that balance *on average*. Unbalanced assignments are possible, potentially creating over-estimates (when comparing a un-tutored, poorer group and a tutored, wealthier group) and under-estimates (when comparing a tutored, poorer group and untutored wealthier group). On the other hand, if we stratify by family income, randomly assigning only within each income bracket, we eliminate those extremes and can better isolate the effect of tutoring.

Caution: stratified randomization is distinct from stratified *sampling*, which would add another layer of complexity. If targeting the PATE, we are still assuming a simple random sample above.

Figure 5 illustrates of the benefit of stratification in a toy example where two groups of units have small positive or even negative treatment effects and two groups have large positive ones. Here, if we stratify within groups 1-4, we actually get the same $\hat{\tau}$ every time without error (things are not usually so perfect!). Otherwise, we can get extremes such as the ones illustrated, which lead to variable and very incorrect $\hat{\tau}$. This can happen even if the estimator is

still on average correct i.e. unbiased. Note: this also illustrates why, for cluster randomization in Module 4, we do not want homogeneous clusters!

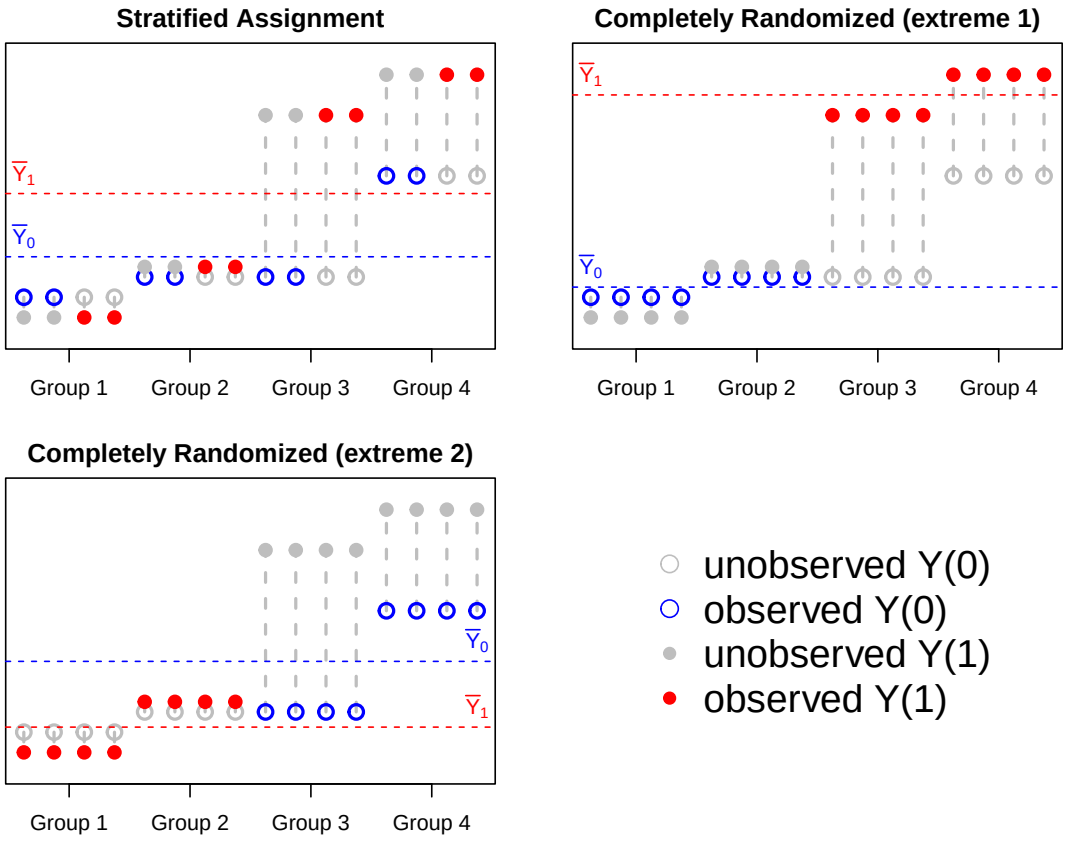


Figure 5: Illustration of the benefit of stratification in a toy example. Each pair of dots vertically connected by a line represents a single unit. The shaded dot represents the units' potential outcome under treatment and the open dot represents the potential outcome under no treatment. Those in gray are unobserved while those in red or blue are observed, with red representing the treatment group and blue the control group.

4 Module 4: Linear Regression for Experiments

Big picture: The goal of this module is to connect linear regression to estimation of average causal effects – still in the experimental context. We do not introduce any new estimands or identification strategies, but regression is such an important and common approach that it is worth studying under what conditions a regression coefficient can be interpreted as a causal effect. In experiments, regressing Y on T ends up algebraically identical to Neyman's ATE estimator because of randomization. We also have to consider the issue of whether the standard regression estimates of uncertainty are valid. Often, these make too-restrictive assumptions, and we need adjusted standard errors.

4.1 Key Concept: (Linear) Structural Equation Model

"Structural Equation Model" (SEM) is somewhat broadly used term which in our course will refer to a model which poses causal relationships instead of only statistical (association-based) ones. In a linear structural equation model, those relationships are linear in the parameters. For example, the simple linear regression model $Y = \alpha + X\beta + \epsilon$ for univariate X and with $\mathbb{E}(\epsilon) = 0$ does **not** pose that β is a causal effect – it is only a statement about **correlation**. In fact, you can show that if $\text{Cov}(\epsilon, X) = 0$ then $\beta = \frac{\text{Cov}(X,Y)}{Var(X)}$ which could in general reflect a non-causal correlation.

If, instead, we pose $Y = \alpha + X\beta + \epsilon$ as a **structural** model, then we are *asserting* X to have average causal effect β on Y. If we then fit a regression and conclude about whether we have evidence $\beta \neq 0$ with the usual regression statistical tests, our assumption allows us to conclude about the size of the causal effect. Even if we conclude $\beta \neq 0$ with some point estimate $\hat{\beta}$, this is NOT the same **discovering** a causal effect. We **assumed** that any observed correlation between X and Y is a causal effect. If that assumption is wrong, we've only discovered a correlation.

In general, structural equation models are also used when studying observational data, and there we will require other arguments to allow interpreting estimates as causal – essentially, that at least conditionally, we still have some random treatment assignment. Throughout, we will end up *assuming* causal structure (hopefully with some justification!). There is another line of work called 'causal discovery' that we will not address in this course.

4.1.1 Basic structural linear models for this module

Slide 3 poses two linear structural equation models for a binary treatment and uses potential outcome notation to make the causal modeling explicit. The first is a constant additive effects model

$$Y_i(t) = \alpha + \beta t + \epsilon_i \quad \text{with} \quad \mathbb{E}(\epsilon_i) = 0$$

where $Y_i(1) - Y_i(0) = \alpha + \beta - \alpha = \beta$ for every unit i. The second model is a heterogeneous treatment effect model

$$Y_i(t) = \alpha + \beta_i t + \epsilon_i \;=\; \alpha + \beta t + (\beta_i - \beta)t + \epsilon_i \quad \text{with} \quad \mathbb{E}(\epsilon_i) = 0 \;\text{ and let }\; \epsilon_i(t) = (\beta_i - \beta)t + \epsilon_i$$

where $\beta = \mathbb{E}(\beta_i) = \mathbb{E}(Y_i(1) - Y_i(0))$. In fact, the two models above end up being indistinguishable from data (Figure 6). As in Module 3, the individual treatment effects β_i are not identifiable – they end up as part of a new error term $\epsilon_i(t) = (\beta - \beta_i)t + \epsilon_i$, which has expectation 0 if $\mathbb{E}(\epsilon_i) = 0$. The ATEs are the same.

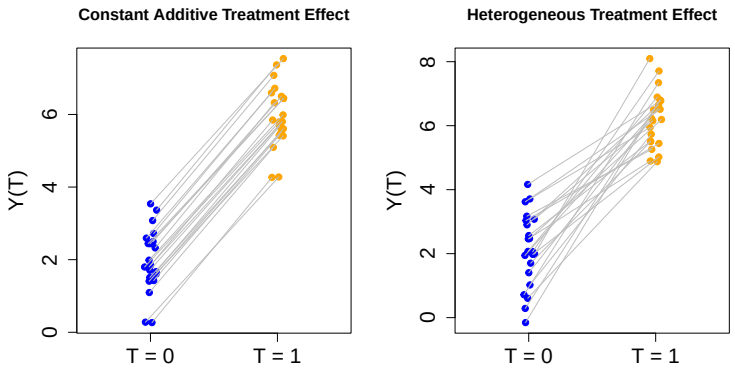


Figure 6: 20 draws of $Y_i(0), Y_i(1)$ (connected by grey line for each unit) according to models on Slide 3 with ϵ_i (left) or $\epsilon_i(t)$ (right) generated from normal distributions with homogeneous variance. Notice how the marginal distributions of treatment and control look the same between the two models.

Issue 1: Where does $\mathbb{E}(\epsilon_i) = 0$ come from? Is that another assumption? As long as we assume SUTVA (so ϵ_i does not actually depend on other units' treatments or on t), this is not a separate assumption. If we set $\mathbb{E}(\epsilon_i) = c$ for any constant c, we could always let $\alpha^* = \alpha + c$ and $\epsilon_i^* = \epsilon_i - c$ and have $\alpha + \beta t + \epsilon = \alpha^* + \beta t + \epsilon_i^*$ and use the later as our model with $\mathbb{E}(\epsilon_i^*) = 0$.

Issue 2: What assumptions ARE we making then? The above discussion only involves defining structural models. The need for assumptions arises once we want to replace $Y_i(t)$ with Y_i and argue that correlations are causal. Without this, in general, confounding could lead to selection of observations with certain kinds of ϵ_i (and certain kinds of $\beta_i - \beta$ for the second model) into the treat and control group, yielding $\mathbb{E}(\epsilon_i|T_i = t) \neq 0$ and $\mathbb{E}(\epsilon_i(t)|T_i = t) \neq 0$. That is, the key assumption we need for identification is **exogeneity** $\mathbb{E}(\epsilon_i|T) = 0$.[6] Note that this also implies $\mathbb{E}(\epsilon_i(t)|T_i = t) = 0$ (Exercise: verify this). Exogeneity thus implies for both models that

$$\mathbb{E}(Y_i|T_i = t) = \mathbb{E}(Y_i(t)|T_i = t) = \alpha + \beta t$$

which means

$$\mathbb{E}(Y_i|T_i = 1) - \mathbb{E}(Y_i|T_i = 0) = \beta$$

the left side is something we can estimate in an unbiased way from data, since it only involves observables. Hence we have **identification** of ATE, which is β (regardless of whether constant effects or heterogeneous effects model is true).

Issue 3: How does this connect to Neyman? Given we have a randomized experiment (and for PATE, simple random sample) exogeneity is not actually some further assumption we have to make here! It follows from exactly the assumptions we used to justify Neyman's estimators in Module 3. Formally, without any reference to linear models, randomization already implies that

$$\mathbb{E}(Y_i(t)) = \mathbb{E}(Y_i(t)|T_i = t) = \mathbb{E}(Y_i|T_i = t)$$

we could stop there and use the estimator's in Module 3. But if we then write out these quantities in terms of our regression models, this implies

$$\alpha + \beta t = \alpha + \beta t + \mathbb{E}(\epsilon_i|T_i = t)$$

(could also replace with $\mathbb{E}(\epsilon_i(t)|T_i = t)$ for the second model). This implies that we must have $\mathbb{E}(\epsilon_i|T_i = t) = 0$, which is exogeneity. Conversely, suppose we only assume exogeneity. Then the identification argument above shows we can identify β exactly in terms of the $\mathbb{E}(Y_i|T_i = 1) - \mathbb{E}(Y_i|T_i = 0)$ that we got for Neyman. Moreover, exogeneity essentially implies we have treatment randomization because it says that treatment is uncorrelated with any of the other factors out there creating variation in outcomes $Y_i(t)$. Overall, the two (randomization and exogeneity) are two different ways of talking about the same thing here.

[6] Technically this is the full T vector but if we assume no interference, random sampling, ϵ_i is independent of the other treatments. In what follows I write $\mathbb{E}(\epsilon_i|T_i = t)$

4.2 Neyman's Estimators (Module 3) vs Regression Estimator

	Neyman	**Regression**
Model	$\tau = \mathbb{E}(Y_i(1) - Y_i(0))$ is causal effect, might or might not assume constant individual-level effects, no other modelling assumptions	Assume one of two structural linear models: (1) Constant effect $Y_i(t) = \alpha + \beta_i t + \epsilon_i$ or (2) heterogeneous effect $Y_i(t) = \alpha + \beta t + \epsilon_i(t)$ with $\mathbb{E}(\epsilon_i) = 0$, $\mathbb{E}(\beta_i) = \beta$, $\epsilon_i(t) = (\beta_i - \beta)t + \epsilon_i$. Cannot distinguish which is true.
Assumptions	complete randomization, i.i.d. sample for PATE	
Identification	randomization implies $\mathbb{E}(Y_i(t)) = \mathbb{E}(Y_i(t)\|T_i = t) = \mathbb{E}(Y_i\|T_i = t)$ meaning $\mathbb{E}(Y_i(1) - Y_i(0)) = \mathbb{E}(Y_i\|T_i = 1) - \mathbb{E}(Y_i\|T_i = 0)$ and the right side is estimable from observed data using group means.	randomization implies exogeneity $\mathbb{E}(\epsilon_i\|T) = 0$ and $\mathbb{E}(\epsilon_i(t)\|T) = 0$ which means regression on observables unbiased for $\beta = \mathbb{E}(Y_i(1) - Y_i(0))$
Unbiased Estimator for ATE	Turns out they're the same! $\hat{\tau} = \hat{\beta}$	
Variance	PATE variance estimator is unbiased $\frac{\hat{\sigma_1}^2}{n_1} + \frac{\hat{\sigma_0}^2}{n_0}$	If assume homoskedasticity, the standard linear regression variance estimator is not the same as Neyman's. If allow heteroskedasticity between treat and control group, heteroskedasticity-robust EHW estimator is consistent for the variance and, heteroskedasticity-robust HC2 estimator is numerically identical to Neyman's variance estimator. (See Module 4 Review Questions for bias derivation for standard regression estimator)

4.3 Further Resources

1. See Appendix B
2. Textbooks on linear regression include: Agresti (2015) (Chapter 2 mainly relevant for this course, very statistical, graduate level) and Rawlings (1998) (Chapters 1-4 and 9 most relevant, especially 3, this book much more accessible). A good online resource is `https://mattblackwell.github.io/gov2002-book/06_linear_model.html` (Chapter 5-7, with set-up and notation that nicely connects to causal inference). There are multitudes of other textbooks as well as videos on YouTube if you need a refresher (e.g., this one gives a good overview of standard OLS assumptions `https://www.youtube.com/watch?v=a1ntCyeoJ0k`).
3. Imbens and Rubin (2015) Chapter 7

4.4 Extension: Regression and Cluster Randomization

Regression can also be connected to estimating treatment effects for more complicated randomization schemes. In the Module 4 problem set, you will see this for cluster randomization.

4.4.1 Why Cluster?

In cluster randomized designs, treatments are randomly assigned only at the level of groups of units (clusters). This is not an efficient strategy in that it inflates the variance but can be practically necessary in terms of running the experiment on the ground. It can also be away to avoid issues with interference and spillover effects if we suspect there would be interference within clusters if only some units within clusters were treated.[7] We do still assume no interference across clusters. You might ask why we do not simply treat the clusters as our units and apply our usual experimental calculations. In fact, we could do that if we had some cluster-level outcome variable $Y_j(t)$. Often, however, cluster-level outcomes will be aggregates of outcomes for individuals within a cluster and we may really be interested in an average individual-level effect. That is, in cluster randomized experiments, our estimands may still units within clusters (e.g., students within a school). Given $Y_{ij}(t)$, the potential outcome for individual i in cluster j, and given m_j units per cluster, our estimand is still the individual-level SATE (or PATE):

$$SATE = \frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{m_j} Y_{ij}(1) - Y_{ij}((0)$$

That said, our *estimators* will still end up aggregating outcomes over clusters (possibly with some weighting).

4.4.2 What makes a good cluster?

Recall that in stratified randomized design (Section 3.6), we want strata to be internally homogeneous (to allow good treat vs control comparison within strata) and variable across strata (account for other sources of variation). When doing cluster-randomization, the opposite is preferable. Is is better for clusters to be not systematically different from each other and internally variable, so that assigning by cluster is not so different from assigning individually. Intuitively, systematic between-cluster variation makes it harder to detect whether differences in outcome are attributable to treatment assignment or cluster variation. Low variation within clusters is wasteful because we are assigning many similar units only to treatment or only to control without getting to compare them. Formally, we say we want a low intra-class correlation.

Example: in the tutoring experiment, suppose we have children from multiple schools. It might be logistically hard or unfair to only tutor some students from each school, so we might have to settle for assigning *schools* to have a tutoring program or not. Suppose we do still measure student-level outcomes. If the schools are drastically different – e.g., some in very wealthy areas and others in very poor areas – then our observed treatment effect might vary a lot with which schools happen to be assigned to treatment or not. If, on the opposite extreme, we imagine students were randomly assigned to schools and then schools were randomly assigned treatment, then we'd actually be back at individual-level randomization. Even if students are not randomly assigned to schools, if schools contain similar mixes of students, the observed treatment effect will vary less with which schools are treated or not.

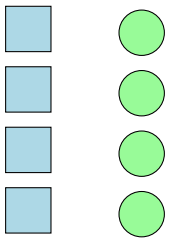


Figure 7: Heuristic diagram: Imagine each row is a cluster. These are ideal clusters. If we assign half the clusters to treatment and half to control, we have equal proportions of each shape in treat and control. Imagine each column is a cluster. This is bad. Now we assign all of one shape to treatment and all of another to control!

[7] It is not that cluster randomization removes the presence of spillover effects. It is that if only everyone or no one within a cluster gets treatment, then we do not allow spillover to be revealed. When we cluster randomize, for example, we do not consider what the treatment effect would be if we only assigned half of the units in each cluster to treatment. One extension that does allow us to get at such effects is called two-stage randomization, where we first randomly select clusters for treatment and then, within treated clusters, randomly assign p% of units in the cluster to be treated for varying values of p. We can then study how treatment effects depend on how much of a cluster was treated.

4.4.3 Estimators:

Given we have randomization at the cluster level, we could calculate our usual Neyman estimator for a cluster-level outcome such as the mean outcome.

$$\hat{\tau}_1 = \frac{1}{J_1}\sum_{j=1}^{J} T_j \bar{Y}_j(1) - \frac{1}{J_1}\sum_{j=1}^{J}(1-T_j)\bar{Y}_j(0) = \frac{1}{J_1}\sum_{j=1}^{J} T_j \sum_{i=1}^{m_j}\frac{Y_{ij}(1)}{m_j} - \frac{1}{J_1}\sum_{j=1}^{J}(1-T_j)\sum_{i=1}^{m_j}\frac{Y_{ij}(0)}{m_j}$$

This is an unbiased estimator for the true average treatment effect on cluster means $\frac{1}{J}\sum_{j=1}^{J}\overline{Y_j(1)} - \overline{Y_j(0)}$ with $\overline{Y_j(t)} = \frac{1}{m_j}\sum_{i=1}^{m_j} Y_{ij}(t)$ but not in general for the SATE because

$$\mathbb{E}(\hat{\tau}_1|\mathcal{O}_n) = \frac{1}{J}\sum_{j=1}^{J}\sum_{i=1}^{m_j}\frac{(Y_{ij}(1)-Y_{ij}(0))}{m_j}$$

If all cluster sizes are the same so that $m_j = m$ and we get a $Jm = n$ in front, then this is unbiased for the SATE. If cluster sizes are similar, this should not be very biased but if, for example, one cluster is much larger than the rest, the SATE and the ATE on cluster means can be quite different. What if, instead, we naively completely ignore clustering and regress Y on T to obtain a point estimate of the ATE? As argued at the start of this module, this is

$$\hat{\tau}_2 = \frac{1}{n_1}\sum_{j=1}^{J}\sum_{i=1}^{m_j} Y_{ij}T_j - \frac{1}{n_0}\sum_{j=1}^{J}\sum_{i=1}^{m_j} Y_{ij}(1-T_j)$$

This is in general biased for the SATE unless all clusters have the same size, in which case it is identical to $\hat{\tau}_1$.[8] An unbiased alternative for the SATE (see Module 4 problem set) is:

$$\hat{\tau}_3 = \frac{J}{nJ_1}\sum_{j=1}^{J}\sum_{i=1}^{m_j} Y_{ij}(1) - \frac{J}{n_0 J_0}\sum_{j=1}^{J}\sum_{i=1}^{m_j} Y_{ij}(0)$$

This is unbiased regardless of whether the clusters have the same size or not! Overall, in the case where all clusters are the same size, the Neyman estimator for effect on cluster-level means, the naive regression, and the third weighted version are identical $\hat{\tau}_1 = \hat{\tau}_2 = \hat{\tau}_3$ but otherwise, they are at least slightly different. Both $\hat{\tau}_1$ and $\hat{\tau}_3$ account for clustering in some sense, but they are motivated by different estimands.

4.4.4 Variance

Assuming equal cluster sizes, the conservative Neyman variance is (from $\hat{\tau}_1$ above) again

$$\mathbb{V}(\hat{\tau}_1|\mathcal{O}_n) \leq \frac{\mathbb{V}(\overline{Y_j(1)})}{J_1} + \frac{\mathbb{V}(\overline{Y_j(0)})}{J_0} \tag{1}$$

which can be estimated with the observed variances of cluster means in the treat and control group. See Module 4 problem set for the adjusted version of the variance for $\hat{\tau}_3$ that works even for unequal cluster sizes. For $\hat{\tau}_2$, we have the advantage that we can consider standard error estimators from regression. However, as in the basic regression case, even if the 'naive' regression is unbiased or yields small bias (cluster sizes not too different), the standard errors are where things become especially tricky. Given clustering, we may not only have violations of homoskedasticity between the treatment and control group; we may also have violations of homoskedasticity across clusters and violations of independence between individual observations. Clustering means that Y_{ij}'s within the same cluster may be more correlated than Y_{ij}'s from different ones, creating non-zero off-diagonal elements in the variance-covariance matrix of Y.

As shown in the lectures slides, **the naive variance estimator** that pretends we have no clustering and only allows for heteroskedasticity in between treatment and control groups (i.e., the same thing we did in the no-clustering case) will under-estimate the variance in the clustering case. In particular, the lecture slides show that:

[8] Finding the expectation is a little tricky because n_1 and n_0 are now also random. However, in the case where all clusters have the same size, $n_1 = J_1 m$ and $n_0 = J_0 m$ are fixed. In the more general case, under some assumptions such as that cluster sizes are drawn i.i.d. from some distribution with $\mathbb{E}(m_j) = m$ with sizes independent of treatment and fixed treatment proportion $\frac{J_1}{J} = k$, then we should be able to make an asymptotic argument as $J \to \infty$ that this estimator is at least consistent for the PATE (i.e., argue $\frac{n_1}{J} = \frac{1}{J}\sum_{j=1}^{J} m_j T_j \to km$ and argue numerator converges to $km\mathbb{E}(Y_i j(1) - Y_i j(0))$.

$$\frac{\mathbb{V}(\overline{Y_j(t)})}{J_t} = \frac{\mathbb{V}(Y_{ij}(t))}{J_t m}(1 + (m-1)\rho_t)$$

where ρ_t is the intraclass correlation (ICC). Think of the variance term on the right side as the equivalent of $\frac{\sigma_t^2}{n_t}$ from the no-clustering case while the left side is the equivalent of this quantity when the outcomes are taken as the cluster means instead (as in $\hat{\tau}_1$ above). The $(1 + (m-1)\rho_t)$ is in practice almost always positive and hence an inflation factor reflecting how clustering almost always increases the variance of our estimator. The formula for ρ_t is

$$\rho_t = \frac{\text{Cov}(Y_{ij}(t), Y_{i'j}(t))}{\mathbb{V}(Y_{ij}(t))}$$

If clusters are highly variable internally relative to the total variance in the sample, the numerator will be small and ICC is low. If the clusters are very uniform relative to total variance in the sample, ICC is high (and clustering is very inefficient). See Module 4 problem set question 5 for how to estimate the ICC using regression of Y on cluster so see how much variation in Y is explained by cluster.

Instead of the naive estimator, we should use a **cluster-robust variance estimator**. One from (Liang and Zeger, 1986) is

$$\widehat{V_{\text{cluster}}((\hat{\alpha}, \hat{\beta}) \mid X)} = \left(\sum_{j=1}^{J} X_j^\top X_j\right)^{-1} \left(\sum_{j=1}^{J} X_j^\top r_j r_j^T X_j\right) \left(\sum_{j=1}^{J} X_j^\top X_j\right)^{-1}$$

Where r_j is the vector of residuals for cluster j from the naive regression. This is consistent as $J \to \infty$ for the Neyman variance in Equation 1.[9] The further-improved, bias-adjusted CRS estimator (Bell and McCaffrey, 2002) again (as in the HC2 estimator above) uses the projection matrix H to further improve the estimator, replacing the center-piece (the 'meat' of the sandwich) with

$$\text{meat} = \sum_{j=1}^{J} X_j^T (I_{m_j} - H_j)^{-1/2} r_j r_j^T (I_{m_j} - H_j)^{-1/2} X_j$$

where $X_j = (1_{m_j}^T, T_j^T)^T$ $H_j = X_j(X^T X)^{-1} X_j$ and r_j contains the residuals for cluster j. Intuitively, by including the residual vector outer product $r_j r_j^T$ for a cluster in the estimator, these estimators account for within-cluster correlation. **An overall lesson: cluster the standard errors at the level of treatment assignment.**

[9] For the non-equal cluster size case, you could use an adjusted version based on Module 4 Problem Set Question 2, using what the solutions call $\tilde{Y}_i$ instead of Y_i.

4.5 Other Regression Extensions

4.5.1 Adding more covariates (regression adjustment)

If we have a completely randomized experiment and hence did not do stratified randomization but do have access to **pre-treatment covariates,** it can sometimes be beneficial to add them into the regression (this is called regression adjustment). That is, we might have:

$$Y_i = \alpha + \beta T_i + \gamma^T \tilde{X}_i + \epsilon_i$$

where $\tilde{X}_i$ is a mean-centered covariate ($\tilde{X}_i = X_i - \bar{X}_n$). You can read more about this in (Imbens and Rubin, 2015, Ch. 7), but a few main points:

- In the model above, the focus is still on estimating β, not on γ, which are called **nuisance parameters** (unknowns only of interest insofar as we need to estimate them to estimate β)
- The OLS estimator is now only asymptotically unbiased (**consistent**) from a super-population perspective as $n \to \infty$ but this holds regardless of the distribution of $\tilde{X}$ and even if the model is not well-specified (e.g. no true linear relationship to X)
- The potential benefit is improved **efficiency** - i.e. lower variance - if X accounts for some of the variation in Y. Here, model specification *does* matter. If the model is very mis-specified, the variance could get worse. That said, in many practical cases, for regressions with reasonable looking diagnostics (e.g. no huge outliers, no variable that should be on log scale included without log scale,...), if X accounts for some of the variation in Y, there is an efficiency gain.
- It is crucial to NOT regress on **post-treatment covariates** that were measured after treatment and could be affected by treatment. Consistency results depend on treatment assignment T being independent of X (because of randomization)

Example: in the tutoring experiment with math score as outcome, suppose we did not stratify by parent income, but we do suspect it is correlated with the outcome. We can regress outcome on treatment and parent income to isolate the treatment effect *with parent income held constant* ('controlled for'). In small samples, because we did not stratify, there's no guarantee of having a balanced number of children from various parental income backgrounds between tutored and untutored groups. Asymptotically, however, that balance should hold. As an example of why not to regress on post-treatment covariates, imagine that students' scores on a math test (our immediate outcome) affect whether they are placed into honors or non-honors algebra in the next year. If we condition on class placement and then look at whether, within those who placed into honors, tutoring had an effect on scores, we'd likely estimate a small effect because everyone who placed into honors scored highly. We could not then detect whether some students only scored high and made it into the honors group because they were tutored. **Caution:** trying many different regressions as above to find one where treatment is statistically significant can lead to problems with multiple testing and p-hacking.

4.5.2 Adding an interaction term

If we add an interaction term into the model, we essentially[10] end up fitting a separate regression for each treatment level:

$$Y_i = \alpha + \beta T_i + \gamma^T \tilde{X}_i + \delta^T T_i \tilde{X}_i + \epsilon_i$$

This model essentially fits two regressions of Y on $\tilde{X}$, one for the treatment group and one for the control group. This could again increase precision, but in the context of randomized experiments, it is not so important as we do not expect the treatment and control group to be systematically different (on average). In observational studies, it will be more important. One thing to note about this interaction model is that the regression coefficient $\hat{\beta}$ is algebraically equivalent to the following **imputation model**:

$$\hat{\beta} = \frac{1}{n} \sum_{i=1}^{n} T_i(Y_i(1) - \widehat{Y_i(0)}) + (1 - T_i)(\widehat{Y_i(1)} - Y_i(0))$$

[10] If we allow heteroskedastic variance, then this is fully true.

We call it an imputation model because given a treated observation, we are imputing the missing potential outcome under control (and vice versa). Note that this is different from just taking a difference in imputed means:

$$\frac{1}{n}\sum_{i=1}^{n}(\widehat{Y_i(1)} - \widehat{Y_i(0)})$$

Because we still use the observed outcomes, too. A few other things to note about the imputation estimator:

- The fact that the interaction model fits separate regressions for treat and control means this imputation is done without 'double dipping' (i.e. using an estimate based on $Y_i(1)$ to impute $Y_i(0)$). This property will also be relevant in the observational context.
- $\hat{\beta}$ is consistent for the true effect β
- Unlike covariate adjustment without interaction, $\hat{\beta}$ is asymptotically at least as efficient (smaller variance) than the regression without covariates. Note this is asymptotic however and things can be different in finite samples, where models with more parameters are harder to estimate.
- Best practice is still to do stratification before randomization!

See Imbens and Rubin (2015) for more.

4.5.3 Are there similar equivalences for stratified randomization?

Given the above equivalence between Neyman and regressing Y on T alone, a logical question is whether we have something similar for stratified randomization (Section 3.6). Can we just regress the outcome on treatment and a strata fixed effect? This would be nice because regression is a widely used tool that many researchers know how to implement. However, the answer turns out to be no in general (though yes in special cases). If we regress Y on T and a stratum indicator, the resulting estimator $\hat{\beta}$ converges to a weighted average of strata-specific treatment effects.

$$\hat{\beta} \xrightarrow{p} \frac{\sum_{j=1}^{J} w_j k_j (1-k_j)\mathbb{E}(Y_i(1) - Y_i(0))}{\sum_{j=1}^{J} w_j k_j (1-k_j)}$$

where $w_j = \frac{n_j}{n}$ are strata proportions as a fraction of the total sample and $k_j = \frac{n_{j1}}{n_j}$ are treatment proportions within each strata and $k_j(1-k_j)$ is the variance of treatment in each strata.

This is not in general equivalent to Neyman's stratified randomization estimator and hence not in general consistent for the true ATE τ. The problem has to do with the fact that stratified randomization may have different treatment proportions in different strata and the regression will reflect this (up-weighting strata with higher treatment variance $k_j(1-k_j)$). If the assignment probability is constant across strata[11] or if the treatment effect is identical across strata, then we do have equivalence between the regression estimator and Neyman's estimator. See (Imbens and Rubin, 2015, Ch 9.) for more.

Note 1: In the previous section, we noted that $\hat{\beta}$ is consistent in a regression where treatment is independent of additional pre-treatment covariate X. What if we treat strata as such an additional covaritate? Above, if k_j are all the same, then treatment is independent of the strata dummy covariates and we do have consistency. But if the proportion varies by strata, then we no longer have independence and lose consistency. So the stratification result is not a contradiction of the covariate adjustment result.

Note 2: The stratification result actually gives some broader insight. As we will see in Module 7 in the observational context, a common required assumption is conditional unconfoundedness $Y(1), Y(0) \perp T|X$ for covariates X. Here, think of strata as X. Under stratified randomization, this unconfoundedness holds. But the resulting regression *will* involve some weighting that up-weights those with high variance in treatment and downweights those with lower variance in treatment (probability of treatment closer to 0 and 1)– this happens more broadly.

[11] This is trivially true in matched pairs experiments, where it is always 1/2

5 Module 5: Instrumental Variables (IV)

Big Picture: if treatment is not randomized, but we have an *instrument* for treatment, which we can think of as a *randomized encouragement* to treat, then we can still identify the average treatment effect, albeit only for a special sub-population called the 'compliers'. In an experimental setting, IV helps us deal with non-compliance. In observational settings, we need to argue carefully that certain assumptions hold about our 'encouragement' variable.

5.1 Motivation in terms of compliance and encouragement to treat

Suppose in the tutoring experiment, we cannot force students to enroll in tutoring. Instead, we can only randomize whether we *offer* them the tutoring or whether we *advertise* the tutoring to them. Some may not take the offer (non-compliance) and if it is possible to enroll without an offer, some may enroll without tutoring being advertised to them (also called non-compliance). In this context, we could take one of at least three approaches:

Strategy	**Assumption**	**Identifies**	**Drawback**
Ignore non-compliance and do usual calculation	valid if non-compliance is at random	ATE of original treatment	usually unrealistic
Intention to treat effect - do usual calculations with "encouragement to get treatment" as the treatment variable	usual randomized experiment assumptions, but for new treatment variable	average intention to treat effect (ITT) of different but related treatment for whole population	do not get ATE for treatment of direct interest
Instrumental variables	see below	ATE only for a sub-population (compliers)	do not get ATE for full population of interest

The intuition for IV is that if offering people tutoring causes some people to be tutored (meaning they would not have enrolled if not offered) and we see a difference in outcomes between those offered tutoring and not offered tutoring, some of that difference could reflect a treatment effect. A complication is that there could also be people who would have enrolled (or not enrolled) either way. Those people might be systematically different types of people and might also be affecting the final observed difference in outcomes between treated and untreated. Hence we cannot just look at the raw differences between the tutored and untutored groups without some adjustment.

5.2 Set-up for Binary Instrument

$Z_i \in \{0,1\}$	The instrument (random encouragement)
$T_i(1), T_i(0)$	Potential treatment $T_i(z)$ (two possible)
$T_i = T_i(Z_i)$	Observed treatment
$Y_i(z,t)$	Potential outcome under $T_i = t$ and $Z_i = z$ (four possible)
$Y_i = Y_i(Z_i) = Y_i(Z_i, T_i(Z_i))$	Observed outcome

Potential Treatments: we can sort people into principal treatment strata based on their potential treatments.

Type	$T_i(0), T_i(1)$	
Complier	$(0,1)$	Only enrolls in tutoring if offered
Always-taker	$(1,1)$	Always enrolls
Never-taker	$(0,0)$	Never enrolls
Defier	$(1,0)$	Only enrolled if tutoring not offered

These are key, because we can only identify a **Complier Average Treatment Effect (CATE)** (see below).

Potential Outcomes: For a given unit, we can in theory imagine outcomes under four possible configurations of treatment and instrument: $Y_i(z,t) \in \{Y_i(1,1), Y_i(0,1), Y_i(1,0), Y_i(0,0)\}$. However, for any given unit, only two of these outcomes are actually observable in the sense that some randomizations would produce them. For example, for an always-taker, $Y_i(0,1)$ and $Y_i(1,1)$ are observable but $y_I(0,0)$ and $Y_i(1,0)$ are not. The following table gives, for each principle treatment strata, which hypothetical outcomes are observable.

$Y(z,t)$	Complier	Always-Taker	Never-Taker	Defier
$Y(0,0)$	✓	×	✓	×
$Y(1,0)$	×	×	✓	✓
$Y(0,1)$	×	✓	×	✓
$Y(1,1)$	✓	✓	×	×

Table 2: × = cannot be observed under any encouragement. ✓= can be observed under some encouragement

In the identification strategy below, we sort people by their principal treatment strata, not their values of $Y(z,t)$ (if Y is continuous, this makes little sense anyway). When we consider observed Y_i and how they link to causal quantities, we only ever have $Y_i = Y_i(Z_i, T_i(Z_i))$ – each assignment of Z_i reveals a single 'path' from Z to T to Y. Still, it is good to keep in mind that this hypothetical full set of potential outcomes exists. For example, in the case where the exclusion restriction does not hold (see next section), we could imagine defining a sample average treatment effect as a function of z:

$$ATE(z) = \frac{1}{n}\sum_{i=1}^{n} Y_i(z,1) - Y_i(z,0)$$

That is, we could think of the average effect under encouragement or under no encouragement, and this theoretical quantity would involve all the potential outcomes.

5.3 Core IV Assumptions

In addition to our SUTVA assumptions from Module 1 (no interference, consistency), the following are required for IV identification to be valid:

1. **Exogeneity** (randomized encouragement). E.g., randomly offer tutoring.

   $$Z_i \perp\!\!\!\perp T_i(0), T_i(1), Y_i(0), Y_i(1)$$

   (We do NOT now have $T_i \perp\!\!\!\perp Y_i(0), Y_i(1)$ per se. There could, as in Figure 8 be other factors affecting treatment and outcome, making them dependent.) **Note:** cannot check this with data.

2. **Exclusion** - Instrument only affects outcome via treatment, meaning if treatment is held constant, there should be no relationship between Z and Y – i.e., no direct effect of Z on outcome.

   $$Y_i(1,t) = Y_i(0,t) \qquad \text{for } t = 0, 1$$

   Under this assumption, it makes sense to write $Y_i(t)$ again and speak of units' having a tuple of the form $(Y_i(t), T_i(z))$. In Table 2, the rows essentially collapse to two rows as the potential outcomes in the first two rows and the potential outcomes in the second two rows are always the same.

   Although this assumption technically applies to the hypothetical quantities in Table 2 for all principal strata types, it is practically most relevant for always-takers and never-takers as for these strata, it says that the observable potential outcomes are the same $Y_i(0, T_i(0)) = Y_i(1, T_i(1))$. Note also that we cannot check this assumption with data, since we cannot observe both $Y_i(1, T_i(1))$ and $Y_i(0, T_i(0))$ for any observation and hence do not know who the never- or always-takers are or what all their potential outcomes are. As an example, this would say that being offered tutoring has no affect on math scores, except via its effect on tutoring uptake.

3. **Monotonicity** - there are no defiers. Says that the instrument (encouragement) only ever has no effect or causes people to take the treatment. In tutoring example, means there is no one who will only enroll in tutoring if not offered tutoring. If enrolling requires an offer, this is trivially satisfied. If the offer is more an advertisement, we assume that advertisement does not ever disuade people from enrolling. **Note**: cannot check this with data.

4. **Relevance** Z has an effect on T for at least some units (else, there are no compliers and get division by 0 in identification equality below). **Note**: can check this assumption with data by estimating ATE of instrument on treatment.

5.4 Identification and Estimand

5.4.1 Intuition: Decomposing correlations

A perfect instrument would be one such that everyone is a complier. Then the instrument is essentially identical to the treatment and we have identification by the usual experimental logic. If I have a strong instrument, meaning one that often causes people to take the treatment (many compliers), and if the treatment has an effect on outcome, I'd expect to observe a correlation between instrument and treatment, a correlation between treatment and outcome, and a correlation between instrument and outcome. Instrumental variables identification is about decomposing and arranging those correlations so that they give us an estimate of the causal relationship between T and Y (**correlation is not causation...but the two are sometimes connected!**)

You may sometimes see the exclusion restriction represented with a Directed Acyclic Graph (DAG) like the following. DAGs are another approach to representing causal ideas and we have not covered them yet.

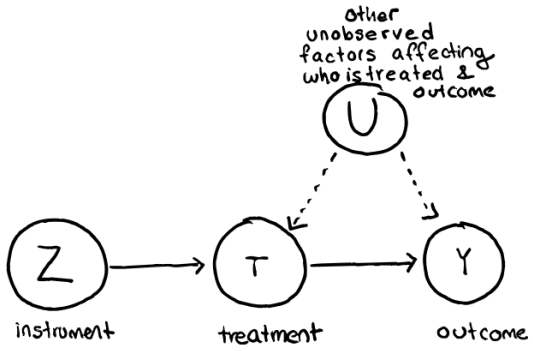


Figure 8: DAG for IV.

For now, just note that the fact that Z is connected to Y via T creates correlation between Z and Y, but what we really want is to isolate the causal connection between T and Y. The fact that there is no 'path' of arrows we can follow from Z to Y without going through T reflects the **exclusion** restriction. There can be other, possibly unobserved, factors U affecting both treatment and outcome with the result, for example, that always-takers and never-takers are systematically different from compliers. Without any adjustment, we can then have non-causal correlations between T and Y. But the idea is that randomizing Z creates an 'exogeneous shock' for a subset of the population (compliers) that breaks the $U \to T$ arrow for that subset. That is, compliers take treatment *because* they were encouraged to, and randomization of Z means that other factors U should balance out on average between the $Z = 0, T = 0$ compliers and the $Z = 1, T = 1$ compliers. By comparison, the $Z = 1, T = 1$ always-takers and $Z = 0, T = 0$ never-takers only appear in one of the two treatment groups and their U do not get balanced between $T = 0, T = 1$. By randomization, they do get balanced between the $Z = 0$ and $Z = 1$ group. Intuitively, exclusion means we can assume that they do not contribute to the correlation between Z and Y. In summary, our assumptions allow the following interpretations of the observational correlations:

- $\text{Cov}(T, Y)$ reflects compliers but also potentially other factors U that might, for example, make the always-takers take treatment and have higher outcomes and the never-takers not take treatment and have lower outcomes
- $\text{Cov}(Z, T)$ positive value reflects presence of compliers – assumed no defiers and by definition, always and never-takers have no effect of Z on T
- $\text{Cov}(Z, Y)$ non-zero value reflects effect of encouragement on outcome but this is driven by compliers – always-takers and never-takers are balanced between $Z = 0, Z = 1$ groups on average so the impact of U is balanced out and by exclusion assumption, have no difference in outcome for $Z = 0$ vs $Z = 1$

We will end up using $Cov(Z, T)$ and $Cov(Z, Y)$ to isolate the effect on compliers.

5.4.2 Formal Identification

Use the Law of Total Expectation to decompose the ATE of the instrument on the outcome (ITT - Intention to Treat Effect) by principal stratum. Let $C = \{i : T_i(0), T_i(1) = (0,1)\}$ and similarly define sets A, N, S for the always-takers, never-takers, and defiers.[12] Then the intention to treat effect on compliers is

$$\text{ITT}_c = \mathbb{E}(Y_i(Z_i = 1) - Y_i(Z_i = 0) | i \in C)$$

Note that for compliers, since $Y_i(Z_i) = Y_i(T_i)$ the ITT_c is the **Complier Average Treatment Effect** (CATE) [13]

$$\text{ITT}_c = CATE = \mathbb{E}(Y_i(T_i = 1) - Y_i(T_i = 0) | i \in C)$$

For always-takers and never-takers, and defiers, we have (noting that these potential outcomes are all $Y(z)$, not $Y(t)$)

$$\begin{aligned}
\text{ITT}_a &= \mathbb{E}(Y_i(1) - Y_i(0) | i \in A) = 0 \qquad \text{(by exclusion assumption)} \\
\text{ITT}_n &= \mathbb{E}(Y_i(1) - Y_i(0) | i \in N) = 0 \qquad \text{(by exclusion assumption)} \\
\text{ITT}_d &= \mathbb{E}(Y_i(1) - Y_i(0) | i \in D)
\end{aligned}$$

By the Law of Total Expectation, (following Slide 5) letting p_s be the probability of principal stratum s.

$$\begin{aligned}
\mathbb{E}(Y_i(1) - \mathbb{E}(Y_i(0)) = \text{ITT} &= \text{ITT}_c p_c + \text{ITT}_d p_d + \text{ITT}_n p_n + \text{ITT}_a p_a \\
&= \text{ITT}_c p_c + \text{ITT}_d * 0 + 0 + 0 \qquad \text{(monotonicity says } p_d = 0) \\
&= \text{ITT}_c p_c
\end{aligned}$$

Slide 6) uses exogeneity and monotonicity to show that $p_c = \mathbb{E}(T_i(1) - T_i(0))$, which is the population average effect of the instrument on treatment and is identifiable via the standard argument of Module 3 given the exogeneity assumption. A related way of arguing this equality is to create a table of which principal treatment strata can be present in each observed (Z, T) combination. For example, if $Z = 0$ and $T = 1$, then this must be an always-taker because we assumed no defiers.

	$Z = 0$	$Z = 1$
$T = 0$	$(0,1), (0,0)$	$(0,0)$
$T = 1$	$(1,1)$	$(0,1), (1,1)$

Table 3: Cell entries are $(T(0), T(1))$ combinations

This table shows us that we can detect the complier group via a kind of 'displacement' in either row. That is, within the $T = 1$ row, the always-takers $(1,1)$ should be randomly split between $Z = 0$ and $Z = 1$ by exogeneity while the compliers will only show up in $Z = 1$, increasing the proportion $P(Z = 1 | T = 1)$ relative to $P(Z = 0 | T = 1)$. We then have $P_c = P(T_i = 1 | Z_i = 1) - P(T_i = 1 | Z_i = 0)$, which exactly corresponds to $\mathbb{E}(T_i(1) | Z_i = 1) - \mathbb{E}(T_i(0) | Z_i = 0) = \mathbb{E}(T_i(1) - T_i(0))$. Note that we could also have written $P_c = P(T_i = 0 | Z_i = 0) - P(T_i = 0 | Z_i = 1)$ and done the same thing in terms of $(1 - T_i(t))$.

The $\text{ITT} = \mathbb{E}(Y_i(1) - Y_i(0))$ is also identifiable by the exogeneity assumption. Hence we have **identification** – we've linked what we want to know to what we can estimate!

$$\text{ITT}_c = \frac{\text{ITT}}{p_c} = \frac{\mathbb{E}(Y_i(1) - Y_i(0))}{\mathbb{E}(T_i(1) - T_i(0))} = \frac{\text{Cov}(Y_i, Z_i)}{\text{Cov}(T_i, Z_i)} \tag{2}$$

[12] These could represent a partition of a finite sample or, more abstractly, a partition of a super-population into the four types. Below we use super-population notation but you could replace it with quantities like $\text{ITT}_Y = \frac{1}{n} \sum_{i \in C} Y_i(1) - Y_i(0)$ and do a finite sample analysis without making claims that effects generalize outside the sample.

[13] **Acronym Note**: CATE also stands for Conditional Average Treatment Effect. The (Complier)ATE is a partucular kind of (Conditional)ATE, which is more generally an ATE defined within some sub-population. We could have a conditional average treatment effect by sex, for example. Note also that the CATE is sometimes also called a LATE (Local Average Treatment Effect) in the literature.

(see Appendix B for why the last equality holds)

We can obtain an unbiased estimator for the numerator and denominator by regressing Y_i on T_i and on Z_i. Since Z_i is binary, this will be a ratio of differences in means. It is also possible to show that this is numerically equivalent to two-stage least squares.

5.5 Two-Stage Least Squares

1. **Stage 1**: Regress T on Z and calculate fitted values $\hat{T}$ (intuition: the part of treatment explained by instrument – the exogenous part coming from exogeneous variation in instrument)

$$T_i = \alpha + Z_i\beta + \epsilon_i$$

$$\hat{T}_i = \hat{\alpha} + Z_i\hat{\beta}$$

2. **Stage 2**: Regress Y on $\hat{T}$ (intuition: the variation in Y explained by exogeneous variation in T is the treatment effect)

$$Y_i = \gamma + \hat{T}_i\delta + \eta_i$$

   $\hat{\delta}$ is then equivalent to Wald Estimator.

Note that by the same argument in module 4 (Covariate adjustment), you can also add pre-encouragement covariates X into these regressions for a possible efficiency gain.

5.6 Summary

To parallel the table in Section 3.2, we now have

Estimand	CATE (really, we'd like ATE but cannot identify it)
Identification Strategy	Arguing we have an instrument for which assumptions in Section 5.3 hold so that get equation (2)
Estimator	Wald Estimator: $$\hat{IV}_{wald} = \frac{\widehat{\text{ITT}}_Y}{\widehat{\text{ITT}}_T}$$ Algebraically equivalent to two-stage least squares. By consistency of the numerator and denominator and Slutsky, this is consistent for (converges in probability to) the CATE. It is also **asymptotically normal** (apply multivariate CLT to numerator and denominator and then the Delta Method)
Asymptotic Variance of Estimator (Derived in Module 5 review sheet)	Let $\rho = \text{Cov}(\widehat{\text{ITT}}_Y, \widehat{\text{ITT}}_T)$ $$\mathbb{V}(\hat{IV}_{wald}) \approx \frac{1}{(\text{ITT}_T)^4}\left(\text{ITT}_T^2\mathbb{V}(\widehat{\text{ITT}}_T) + \text{ITT}_Y^2\mathbb{V}(\hat{\text{ITT}}_T) - 2\text{ITT}_Y\text{ITT}_T\rho\right)$$ *(larger covariance corresponds to stronger instrument and smaller variance)*
Estimate of Asymptotic Variance	Plug in $\widehat{\text{ITT}}_Y$, $\widehat{\text{ITT}}_T$, their variance estimates, and $\hat{\rho}$
Confidence Interval	Asymptotic normality $$\sqrt{n}(\hat{IV}_{wald} - CATE) \xrightarrow{d} N(0, \mathbb{V}(\hat{IV}_{wald}))$$ justifies $(1-\alpha)$ confidence intervals of the form $\hat{IV}_{wald} \pm z_{1-\alpha/2}\sqrt{\hat{\mathbb{V}}(\hat{IV}_{wald})}$.
where...	• See Section 3.2 for difference in means form of $\widehat{\text{ITT}}_Y$ and $\widehat{\text{ITT}}_T$ as well as forms of $\mathbb{V}(\widehat{\text{ITT}}_Y)$, and $\mathbb{V}(\widehat{\text{ITT}}_T)$ and their unbiased estimators. • As derived in Module 5 Review Question 1: $$\rho = \frac{\text{Cov}(Y_i(1), T_i(1))}{n_1} + \frac{\text{Cov}(Y_i(0), T_i(0))}{n_0}$$ • $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of a standard normal

Note: see Module 5 Review Question 3 for formulas for the bias of the Wald estimator when some assumptions do not hold.

5.7 Further notes

1. **IV will not fix your ITT= 0**: Notice that the identification equality and hence the IV estimator is just a re-scaled version of the ITT: $ITT_c = \frac{ITT}{p_c}$. Since $p_c \in [0, 1]$, the estimated CATE will always be greater than or equal to the ITT. If the (estimated) ITT is 0, the (estimated) CATE is, too. If the confidence interval does not rule out ITT being 0, we should not rule out the CATE being so. Hence, doing an IV analysis is not a way to get around the issue that your ITT estimate is not significant!

2. **IV in Observational Setting**: Encouragement to treat and non-compliance are useful for thinking about what IV is in general, so we frame IV this way. However, with IV, we are moving past the strictly experimental setting. More generally, we just need to argue we have a variable that meets the conditions to be an instrument below (easier said than done!). The instrument does ultimately have to have some random variation in the sense that it needs to meet the exogeneity assumption – but sometimes we might discover this in a real-world setting without ourselves doing the randomization (e.g., a policy allocated by lottery, the weather as arbitrary variation that affects a treatment without directly affecting outcome). Especially in these observational settings, the choice and validity of instruments can be hotly debated and some are critical of instrumental variables methods.

3. **IV and External Validity**: even in a standard experiment, the issue of external validity can be tricky (does effect τ on college students in Fall 2015 apply to retirees in 2023?). IV narrows the scope even more since we only estimate the CATE. Essentially, we pay the price of making more assumptions and getting less external validity for the gain of internal validity (correctly identifying the CATE). But internal validity for who? The compliers are a latent group, and we do not even get to know exactly who is in it. Ideally, we'd like a strong instrument that comes close to being a proxy for treatment itself and yields a large complier group. However, a candidate for a strong instrument may also be at more risk of violating the exclusion assumption (could itself have a separate effect on the outcome). This is the **weak instruments** dilemma - you want something subtle but not something so weak that it creates an extremely tiny complier group. Note also that different choices of valid instruments could yield different kinds of complier populations – there is not one CATE but many possible ones. As we often say in statistics: **there's no free lunch!**

5.8 Further Resources

1. Causal Inference Bootcamp Videos `https://mattmasten.github.io/bootcamp/` See video 2.4 for motivation in terms of noncompliance, section 4 for an extended, not-very-mathematical overview of IV ideas

2. Textbook with Chapter on IV that includes code examples using `AER` package and `ivreg` function `https://www.econometrics-with-r.org/12-ivr.html`

3. Another nice chapter on IV in Cunningham (2021):`https://mixtape.scunning.com/07-instrumental_variables`. See also this chapter for a section on **weak instruments** and further discussion of the schooling example used in the Govt problem set for Module 5

4. Imbens and Rubin (2015) Chapter 23-24

5. Feller et al. (2017) contains some nice examples of non-compliance and a brief overview of the principal stratum perspective on this.

5.9 Extension: IV with multi-valued treatment

In this setting, we still have a binary instrument (encouraged or not) that may encourage different degrees of uptake. In a tutoring experiment, children may attend tutoring with different frequencies. In a medical experiment, people make take different doses of medicine or tale the medicine for different lengths of time. Formally, we have:

- **Instrument** $Z_i \in \{0, 1\}$ as before
- **Treatment** $T_i \in \{0, ..., K-1\}$ (K possible values, numeric or possibly ordinal - assume $T_i \geq T_j$ is meaningful)
- **Outcome** Y_i as before

And the assumptions become

- **Exogeneity** - randomized instrument – as before
- **Exclusion** - no effect on outcome among those for whom instrument has no effect on treatment $Y_i(1, t) = Y_i(0, t)$ for $t = 0, 1, ..., K-1$
- **Monotonicity** - instrument only ever increases treatment (e.g., encourages more tutoring attendance): $T_i(1) \geq T_i(0)$
- **Relevance** - instrument has causal effect on treatment for at least some, ideally many, individuals

Note that there are now many more principal strata than before! There are the people who would have had no treatment ($T_i(0) = 0$) if not encouraged and $T_i(1) = 1$ if encouraged, the people who would have had no treatment if not encouraged and $T_i(1) = 2$ if encouraged, and so on...but also the people who would have $T_i(0) = 1$ if not encouraged and $T_i(1) = 2$ if encouraged! Monotonicity does remove some possibilities, but we are still left with $\frac{K(K+1)}{2}$ possibilities! These include K non-complier strata of people with the same treatment regardless of instrument and $\frac{K(K-1)}{2}$ compliers where treatment is higher under encouragement. **We cannot in general identify all the principal strata proportions, even under monotonicity.**

What CAN we do?

1. **Identify marginal distributions** $P(T_i(z) = k)$. For example, under complete randomization of the instrument:

$$\Pr(T_i(1) = k) = \Pr(T_i(1) = k | Z_i = 1) = \Pr(T_i = k | Z_i = 1)$$

   suggesting the estimator $\frac{1}{n_1} \sum_{i=1}^{n} Z_i I(T_i = k)$.

2. **Dichotomize**: we might define a treatment indicator $W_i = I(T_i > 0)$ and do our usual binary treatment IV analysis for the effect of *any* adoption of treatment. The downside here is we destroy information and might miss dynamics such as larger treatment leading to larger outcome. Moreover, dichotomizing at other values, such as defining $W_i = I(T_i > k)$ can lead to bias under some conditions. You will explore this in Stat 286 Module 5 Question 2 and 4.

3. **2SLS Estimator for average incremental effect**: under some assumptions, we can use 2SLS obtain a consistent estimator of a weighted average increase (decrease) in Y per 1-unit increase in treatment. Recall that 2SLS is algebreically equivalent to the Wald Estimator, which has the form $\frac{\widehat{\text{ITT}_Y}}{\widehat{\text{ITT}_T}}$. To understand what that ratio now represents, we'll use the **Law of Total Expectation** with conditioning on principal strata to rewrite $\frac{\text{ITT}_Y}{\text{ITT}_T}$.

$$\text{ITT}_Y = \mathbb{E}(Y_i(1) - Y_i(0)) = \sum_{k=0}^{K-1}\sum_{j=k+1}^{K-1} \mathbb{E}(Y_i(1) - Y_i(0) \mid T_i(1) = j, T_i(0) = k) \Pr(T_i(1) = j, T_i(0) = k)$$

$$\begin{aligned}\text{ITT}_T = \mathbb{E}(T_i(1) - T_i(0)) &= \sum_{k=0}^{K-1}\sum_{j=k+1}^{K-1} \mathbb{E}(T_i(1) - T_i(0) \mid T_i(1) = j, T_i(0) = k) \Pr(T_i(1) = j, T_i(0) = k) \\ &= \sum_{k=0}^{K-1}\sum_{j=k+1}^{K-1} (j-k) Pr(T_i(1) = j, T_i(0) = k)\end{aligned}$$

The indexing reflects the monotonicity assumption (we only condition on principal strata allowed to exist) and the exclusion assumption, which means that for $T_i(0) = T_i(1) = k$ cases, $Y_i(1) - Y_i(0) = 0$ and the term drops out. For ITT_T, note that for each principal stratum, $T_i(1) - T_i(0)$ is a fixed, known thing. Denote the principal strata probabilities by $P_{jk} = P(T_i(1) = j, T_i(0) = k)$. Then multiplying by $\frac{j-k}{j-k}$ in the numerator, we have

$$IV_{wald} = \frac{\mathbb{E}(Y_i(1) - Y_i(0))}{\mathbb{E}(T_i(1) - T_i(0))} = \frac{\sum_{k=0}^{K-1}\sum_{j=k+1}^{K-1}(j-k)\mathbb{E}(\frac{Y_i(1)-Y_i(0)}{j-k} \mid T_i(1) = j, T_i(0) = k)P_{jk}}{\sum_{k'=0}^{K-1}\sum_{j'=k'+1}^{K-1}(j'-k')P_{j'k'}}$$

Finally, let's take everything except the part involving Y_i and make it a weight so that we can see the form clearly:

$$IV_{wald} = \frac{\mathbb{E}(Y_i(1) - Y_i(0))}{\mathbb{E}(T_i(1) - T_i(0))} = \sum_{k=0}^{K-1}\sum_{j=k+1}^{K-1} w_{jk} E\left(\frac{Y_i(1) - Y_i(0)}{j-k} \mid T_i(1) = j, T_i(0) = k\right)$$

$$w_{jk} = \frac{(j-k)P_{jk}}{\sum_{k'=0}^{K-1}\sum_{j'=k'+1}^{K-1}(j'-k')P_{j'k'}}$$

We can think of $\frac{Y_i(1)-Y_i(0)}{j-k}$ as a slope. It gives the change in outcome for a $j-k$ unit change in treatment. The final form of IV_{wald} is then a weighted average of the slopes across complier strata types. The weights are proportional to the size of the strata (more weight if P_{jk} large) and the size of the jump in treatment for that strata (more weight if the effect of instrument on treatment, $j-k$, is large).

4. **Further constrain space of possible strata:** the weighted average above is what the Wald estimator represents under only the standard assumptions (monotonicity, exclusion, exogeneity). As noted above, in general, we will not be able to identify P_{jk} to get an understanding of which types of people are present and affecting our estimate. However, under some additional assumptions, the weights may become identifiable. Two possible ones are:

   (a) **One-sided non-compliance:** Non-encouraged ($Z_i = 0$) units never get treatment: $T_i(0) = 0$ for all i, so the only non-compliers are those encouraged who do not get treatment. How many strata do we have then? What do the weights become? Consider how you might then be able to estimate the proportions for each of the principal strata (Stat 286 Module 5 Problem Set Question 1)

   (b) **Bounded instrument effect on treatment:** The encouragement never increases treatment by more than one unit $T_i(1) - T_i(0) = 1$ for all i. In this case, the weights become just a reflection of the proportion of each type:

$$w_j = \frac{P_{j,j+1}}{\sum_{j'=0}^{K-2} P_{j',j'+1}}$$

   And we can identify each $P_{j,j}$ and $P_{j,j+1} = P(T_i(0) = j, T_i(1) = j+1)$. To see this, consider the following table showing the possible principal strata $(T_i(0), T_i(1))$ for each combination of observed Z and T

	$Z=0$	$Z=1$
$T=0$	$(0,1),(0,0)$	$(0,0)$
$T=1$	$(1,2),(1,1)$	$(0,1),(1,1)$
$\vdots$	$\vdots$	$\vdots$
$T=j$	$(j,j+1),(j,j)$	$(j-1,j),(j,j)$

Table 4: Cell entries are $(T(0),T(1))$ combinations

We then have (can replace each of the P's with a $\hat{P}$ and calculate sample proportions as estimators)

- $P_{0,0} = P(T=0|Z=1)$
- $P_{0,1} = P((0,0)\cup(0,1)) - P((0,0)) \;=\; P(T=0|Z=0) - P_{0,0}$
- $P_{1,1} = P((0,1),(1,1)) - P((0,1)) = P(T=1|Z=1) - P_{0,1}$
- $P_{1,2} = P((1,2)\cup(1,1)) - P((1,1)) = P(T=1|Z=0) - P_{1,1}$

$\vdots$

Hence in general, we have the recursive formulas

$$P_{j,j} = P((j,j)\cup(j,j+1)) - P((j-1,j)) = P(T=j|Z=0) - P_{j-1,j}$$

$$\begin{aligned} P_{j,j+1} &= P((j,j+1),(j,j)) - P((j,j)) \\ &= P(T=j|Z=0) - P_{j,j} \end{aligned}$$

Modules 6-10: Observational Data

Leveraging Cut-offs and Parallel Trends, Balancing, Matching, Weighting, Conditioning

6 Module 6: Regression Discontinuity

Big Picture: modules 2-5 relied on a known *randomized* treatment assignment mechanism or encouragement to get treatment. We now move to an observational setting where the treatment assignment mechanism is known but *deterministic*. RD designs require that treatment (or at least encouragement to treatment) follows some deterministic rule. We compare units on both sides of a cut-off, and if there is a jump in outcome, then under some assumptions, we can argue it reflects a treatment effect. This requires **extrapolation** – assuming trends near the cut-off hold. The estimand is limited to a local average treatment effect for units at the cut-off, which may limit external validity.

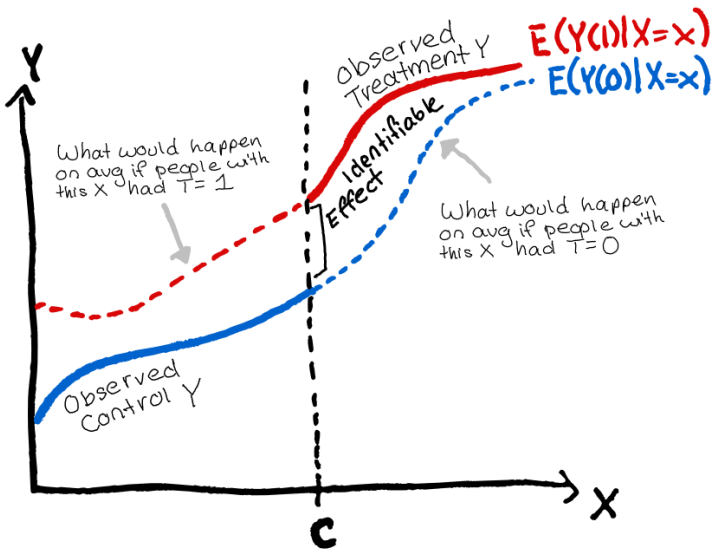


Figure 9: Illustration of the regression discontinuity set-up and assumptions

6.1 Set-up (Binary Treatment, Single Cut-off Case)

- **Binary Treatment** $T_i \in \{0, 1\}$
- **Running variable** X (also sometimes called the forcing variable)
- **Cut-off** c: a value of the running variable which divides those who are treated vs not treated or, in the fuzzy case, less likely to receive treatment vs more likely to receive treatment.

Sharp RD	**Fuzzy RD**
Treatment assigned by deterministic rule: $T_i = I(X_i \geq c)$ or $T_i = I(X_i \leq c)$. Everyone after c is treated.	Non-deterministic treatment assignment - probability of assignment jumps at the cut-off, e.g., because everyone with $X_i \geq c$ is encouraged.
Ex: everyone with pre-test score $\leq c$ is enrolled in tutoring and no one with pre-test score $> c$ is allowed to enroll.	Ex: everyone with pre-test score $\leq c$ is given a voucher for tutoring, though anyone can enroll and those with a voucher may decline to.

- **Continuous-valued outcome** Y_i. Each unit still has potential outcome $Y_i(0)$ and $Y_i(1)$ (occurring on the plot at their X_i value). Instead of just thinking about the unconditional distributions of $Y_i(1)$ and $Y_i(0)$ in the population, we think about the conditional distribution of potential outcomes under treat and control given X. Define

$$f_t(x) = \mathbb{E}(Y_i(t)|X = x)$$

the function describing how the expected value of outcome under treatment t changes with x.

6.2 Estimand and Identification (Sharp RD)

The only estimand we will be able to identify in a RD situation is the effect of treatment τ_c on Y at the threshold.

$$\tau_c = \mathbb{E}(Y_i(1) - Y_i(0)|X = c)$$

This is a Local Average Treatment Effect (LATE).

Assumptions

- **Deterministic treatment assignment**: assume we have a cut-off such that higher values of X are treated ($T_i = I(X_i > c)$).[14] Together with the consistency of potential outcomes assumption, this implies

$$Y_i = Y_i(I(T_i \geq c)) \tag{3}$$

- **Continuity** In a neighborhood of cut-off c, $f_t(x) = \mathbb{E}(Y_i(t)|X = x)$ is continuous. This justifies **extrapolating** from the trend in a region near the cut-off to the other side of the cut-off. The key point is that there should not be anything else going on exactly at $X = c$ that would have caused a jump, even without treatment. It may be that many other factors affect the outcome in general, but at c, there should be no other 'shock.' By definition, continuity implies that for $t = 0$ and $t = 1$

$$\mathbb{E}(Y_i(t)|X = c) = \lim_{x \downarrow c} \mathbb{E}(Y_i(t)|X = x) = \lim_{x \uparrow c} \mathbb{E}(Y_i(t)|X = x) \tag{4}$$

  Continuity is the price we pay for having a deterministic treatment such that there is no **overlap** between treatment and control group. By no overlap, we mean that instead of having units with treatment and control at each X, some X have only treatment and some only control. Because we cannot compare treated and untreated units at the same X, we have to think about extrapolation in a region where we at least have units that are close in their X (i.e. on either side of the cut-off)

- **Functional Form**: we need to assume some continuous functional form(s) for $f_t(x)$ for $t = 0, 1$. This form will define more specifically what we mean when we say $f_t(x)$ is continuous at $x = c$. Note that the key thing is getting a correct functional form within a window around $x = c$ - e.g., some methods only require assuming local linearity around the cut-off.

Identification Argument

Suppose we have deterministic treatment assignment and a correct functional form $f_t(x)$ for $t = 0, 1$. Then we have, at least in principle,[15] observed data which will allow us to estimate the conditional expectation function on either side of the cut-off. If we can form an unbiased (or at least consistent) estimator of the correct functional forms for $T = 0$ and $T = 1$, then under continuity (allows extrapolation) and the consistency of potential outcomes assumption, we have unbiased (consistent) estimators of $\mathbb{E}(Y_i(1)|X = x)$ for $x > c$ and $\mathbb{E}(Y_i(0)|X = x)$ for $x < c$, too.

Formally, suppose we have picked a correct functional form $f_t(x) = g_{\theta_t}(X_i - c)$ where $g_{\theta_t}(0)$ is then the value at the cut-off (this is called **centering** the running variable at the cut-off). If we can obtain an unbiased (or at least consistent) estimator of θ_0, θ_1, e.g., via some regression, then we have identification

$$\mathbb{E}(g_{\hat{\theta}_1}(0) - g_{\hat{\theta}_0}(0)) = \mathbb{E}(Y_i(1) - Y_i(0)|X_i = c) = \tau_c$$

Usually, these models will have the form $f_t(x) = \alpha_t + g_{\theta_t}(X_i - c)$ where α_t is an intercept term, θ_t are nuisance parameters and $g_{\theta_t}(0) = 0$. In that case, the unbiased (or consistent) estimator becomes $\hat{\alpha}_1 - \hat{\alpha}_0$. For example, if the true functional forms are linear $f_t(x) = \alpha_t + \beta_t X_i$ with standard regression assumptions met, then we can use our usual OLS estimators, with the data with $X_i \leq c$ used to form an unbiased estimator of α_0, β_0 and the data with $X_i > c$ used to form an unbiased estimator of α_1, β_1. Note that more generally, equations 4 and **??** imply it would be enough to have a correct functional form in some window around $X = c$ and to be able to correctly estimate that. As we will see below, however, even functional forms further away from c can make a difference for finite sample estimation and especially, variance.

[14] Everything can be flipped or adjusted to the case where units with $X \leq c$ or $X_i > c$ are the treated units.
[15] That is: ignoring finite data issues, taking the infinite data perspective of identification (Section 3.3)

6.3 RD Estimation

Technically, there are infinitely many possible RD estimators – we could in principle fit any functional form we'd like. The simplest is to do linear regression with the data on either side of the cut-off. However, this is not always wise. Below, let $\tilde{X}_i = X_i - c$ throughout so for example $\tilde{X}_i > 0$ is equivalent to $X_i > c$. Also define treatment indicator T_i to reflect whichever side of the cut-off is treatment.

6.3.1 Linear Regression

If we want to pose that $f_t(x) = \alpha_t + \beta_t X_i$, then we can obtain these estimates either by (1) fitting two separate regressions on $\{(X_i, Y_i) : X_i < c\}$ and $\{(X_i, Y_i) : X_i \geq c\}$ or (2) fitting a single regression of Y_i on X_i and T_i with an interaction between treatment indicator and X_i to allow for different slopes and intercepts for each treatment group. These will both yield algebraically identical estimates of $\alpha_0, \beta_0, \alpha_1, \beta_1$. The first is formulated as the solutions to:

$$(\hat{\alpha}_1, \hat{\beta}_1) = \underset{\alpha,\beta}{\arg\min} \sum_{i=1}^{n} T_i(Y_i - \alpha - \beta\tilde{X}_i)^2$$

$$(\hat{\alpha}_0, \hat{\beta}_0) = \underset{\alpha,\beta}{\arg\min} \sum_{i=1}^{n} (1 - T_i)(Y_i - \alpha - \beta\tilde{X}_i)^2$$

We then have $\hat{\tau}_c = \hat{\alpha}_1 - \hat{\alpha}_0$ and, assuming we have an i.i.d. sample and using the fact that these two estimates are based on separate data, we can estimate the **variance** of $\hat{\tau}_c$ using the standard regression estimates of $V(\hat{\alpha}_1) + V(\hat{\alpha}_0)$. If instead, we fit the interaction model, then we have

$$(\hat{\alpha}, \hat{\beta}, \hat{\gamma}, \hat{\delta}) = \underset{\alpha,\beta,\gamma,\delta}{\arg\min} \sum_{i=1}^{n} (Y_i - \alpha - \beta\tilde{X}_i - \gamma T_i - \delta\tilde{X}_i T_i)^2$$

In this case, the fitted line for the $X_i < c$ group is $f(x) = \hat{\alpha} + \hat{\beta}\tilde{x}$ while the fitted line for the $X_i \geq c$ group is $f(x) = (\hat{\alpha} + \hat{\gamma}) + (\hat{\beta} + \hat{\delta})\tilde{x}$. Again, we really care about the $\tilde{x} = 0$ case, which means that our estimator is

$$\hat{\tau}_c = \hat{\alpha} + \hat{\gamma} - \hat{\alpha} = \hat{\gamma}$$

The nice thing about this version is that our estimand is represented by a single parameter in the model and standard R output gives us its variance – albeit under a homoskedasticity assumption. As in Module 4, it would make more sense to use a heteroskedasticity-robust estimator to allow for different variance between treatment and control. Functions like `lm_robust` in the `estimatr` package in `R` implement this.

Notation note: the slides use indexing by + and − to indicate regression of observations above (+) or below (−) the cut-off. In this case, the estimator could be $\hat{\alpha}_+ - \hat{\alpha}_-$ or $\hat{\alpha}_- - \hat{\alpha}_+$ depending on which side corresponds to treatment and control

6.3.2 Polynomial Regression (bad idea)

If we pose a linear model but the actual potential outcome curves are non-linear as in Figure 10, then this can produce biased estimates. One solution might be to fit more flexible models, such as polynomial regression that allows curvature. That is, for the treatment and control group, we could separately fit the following model (or again do one model with interactions)

$$f_t(\tilde{X}_i) = \alpha_t + \delta_{1t}\tilde{X}_i + \delta_{2t}\tilde{X}_i^2 + ... + \delta_{pt}\tilde{X}_i^k$$

for some order k polynomial, and again have

$$\hat{\tau}_c = \hat{\alpha}_1 - \hat{\alpha}_0$$

However, Gelman and Imbens (2019) have argued this is generally a bad idea, showing that it can lead to nonsensical weighting schemes (e.g., up-weighting observations far from the cut-off), that results can be sensitive to polynomial order p, and that inferences can be skewed (e.g., high type I error rate). High degree polynomials are very flexible and can easily overfit the data – at the extreme we might be able interpolate all of the data points. Polynomial fits can also have especially wild behavior near the edges of the data they are fit on - exactly where we want to extrapolate! You will explore some of this bad behavior in the Module 6 Problem Set (question 2).

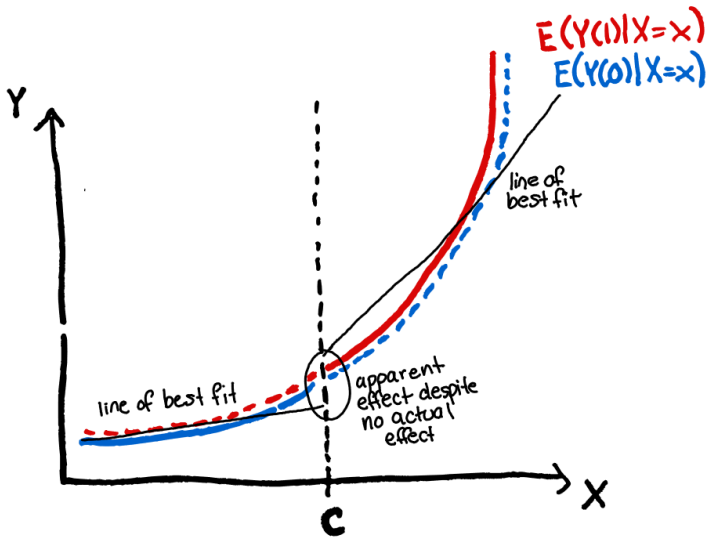


Figure 10: Illustration of incorrect functional form. Notice: if we zoomed in a bit around c, local linearity seems plausible

6.3.3 Local Linear Regression

A better approach is to fit a locally linear model. Most simply, we could imagine throwing away the data outside some window $\tilde{X}_i \in [\Delta, \Delta]$ and fitting a linear regression only close to cut-off c (aka $\tilde{X}_i = 0$), thereby preventing any wild behavior far from $X = c$ from affecting our estimates. We then need to decide on how large Δ should be. More generally, we might not want to discard observations far from $X = c$ entirely but rather down-weight them using a **kernel** function $K(\frac{\tilde{X}_i}{h})$ which, given how far X_i is from c relative to some bandwidth defining the scale (one might call it a tuning parameter), outputs a weight for that observation. We then fit two **kernel regressions**, which are the solutions to

$$(\hat{\alpha}_1, \hat{\beta}_1) = \arg\min_{\alpha,\beta} \sum_{i=1}^{n} T_i(Y_i - \alpha - \beta\tilde{X}_i)^2 K(\frac{\tilde{X}_i}{h}) \tag{5}$$

$$(\hat{\alpha}_0, \hat{\beta}_0) = \arg\min_{\alpha,\beta} \sum_{i=1}^{n} (1 - T_i)(Y_i - \alpha - \beta\tilde{X}_i)^2 K(\frac{\tilde{X}_i}{h}) \tag{6}$$

where again α_1 and α_0 are the expected values at the cut-off and our overall estimator is

$$\hat{\tau}_c = \hat{\alpha}_1 - \hat{\alpha}_0$$

The case where we cut-off at $\pm\Delta$ and then weigh the leftover observations equally is then the special case of $K(u) = I(|u| \leq \Delta)$ with $h = 1$ and $u = \tilde{x} = x - c$. Slide 6 gives some other Kernel options. Note that we could again consider doing a local **polynomial** regression where we add higher order terms to the above expressions. This may give some added flexibility but could still end up being too flexible and even not having a unique solution if k is high and there are few data points near the cut-off. A local first order linear model can be enough.

Choosing a Bandwidth h The choice of bandwidth represents a **bias-variance trade-off**. A too large bandwidth might make non-linearities as described above more of an issue and give too much weight to less relevant observations far from the cut-off. This could create bias. But inherently, as h get small, we throw away more of our data or downweight it more severely, leading to estimates formed on very little data which might be unbiased in repeated sampling but have a high variance because of their sensitivity to particular observations. We might have identification in theory but very poor estimation in practice. Slide 7 gives one approach to navigating this trade-off by minimizing mean squared error. In `R`, the `rdbwselect` function in the `rdrobust` package implements various versions of optimal bandwidth selection for a few different kernel options.

6.4 Local Linear RD as Weighted Regression

How do we fit a local linear regression with a Kernel as described above? While we cannot fit the above regressions using standard OLS estimators, we can use **weighted least squares** with $K(\frac{\tilde{X}_i}{h})$ serving the role of a weight. Consider the of the treated group regression from Equation 5 above. This corresponds exactly to the set-up of a weighted least squares problem where we try to minimize a sum of squares but prioritize fit for observations with larger weights. The general solution in matrix notation for a weighted least squares problem has the form:

$$\hat{\beta} = (X^T W X)^{-1} X^T W Y$$

where

- X is our usual $n \times p$ design matrix of covariates
- Y is our usual $n \times 1$ outcome vector
- W is a diagonal $n \times n$ matrix with weights $w_1, ..., w_n$ on the diagonal and 0's for the rest

Here, for simplicity of notation, suppose there are n_1 treated observations (all above or all below the cut-off) and that we re-index these $1, ..., n_1$. To avoid confusion with the single $\tilde{X}_i$ let us define $n_1 \times 2$ design matrix Z containing in each row the vector $Z_i = (1, \tilde{X}_i)^T$. We let $W = \text{diag}(K(\frac{\tilde{X}_1}{h}), ..., K(\frac{\tilde{X}_{n_1}}{h}))$ for some kernel K and we have $Y = (Y_1, ..., Y_{n_1})^T$. The least squares solution to the above is then

$$(\hat{\alpha_1}, \hat{\beta_1}) = (Z^T W Z)^{-1} Z^T W Y$$

and in particular, letting $e_1 = (1, 0)^T$, we can write

$$\hat{\alpha_1} = e_1^T (Z^T W Z)^{-1} Z^T W Y$$

We could do the same thing for the $T_i = 0$ group with its own (entirely separate) Z, W, Y matrices. In the Module 6 problem set question 2b, you will show that in fact, this can be formulated as a weighted average of the outcome variables under treatment.

$$\hat{\alpha_1} = \sum_{i=1}^{n_1} w_i Y_i \qquad \text{where} \qquad w_i = e_1^T (Z^T W Z)^{-1} Z_i K(\frac{\tilde{X}_i}{h})$$

6.4.1 Deriving the alternative form of this weight on slides

Let $W_i = K(\frac{\tilde{X}_i}{h})$ and note that

$$(Z^T W Z)^{-1} = \begin{bmatrix} \sum_j W_j & \sum_j W_j \tilde{X}_j \\ \sum_j W_j \tilde{X}_j & \sum_j W_j \tilde{X}_j^2 \end{bmatrix}^{-1} = \frac{1}{D} \begin{bmatrix} \sum_j W_j \tilde{X}_j^2 & -\sum_j W_j \tilde{X}_j \\ -\sum_j W_j \tilde{X}_j & \sum_j W_j \end{bmatrix} \text{ with } D = (\sum_j W_j \tilde{X}_j^2)(\sum_i W_i) - (\sum_j W_j \tilde{X}_j)^2$$

and hence

$$(Z^T W Z)^{-1} Z_i W_i = (Z^T W Z)^{-1} (1, X_i)^T W_i = \frac{1}{D} \begin{bmatrix} \sum_j W_j \tilde{X}_j^2 - (\sum_j W_j \tilde{X}_j) \tilde{X}_i \\ -\sum_j W_j \tilde{X}_j + (\sum_j W_j) \tilde{X}_i \end{bmatrix} W_i$$

the multiplication by e_1^T gives only the first element of this matrix, so our weight expression is

$$\begin{aligned} w_i &= \frac{\sum_j W_j \tilde{X}_j^2 - (\sum_j W_j \tilde{X}_j) \tilde{X}_i}{D} * W_i = \frac{\sum_j W_j \tilde{X}_j^2 - (\sum_j W_j \tilde{X}_j) \tilde{X}_i}{(\sum_j W_j \tilde{X}_j^2)(\sum_i W_i) - (\sum_j W_j \tilde{X}_j)^2} * W_i \\ &= \frac{W_i \left(\sum_j W_j \tilde{X}_j^2 - (\sum_j W_j \tilde{X}_j) \tilde{X}_i \right)}{\sum_i W_i \left((\sum_j W_j \tilde{X}_j^2) - (\sum_j W_j \tilde{X}_j) \tilde{X}_i \right)} = \frac{W_i \left(\frac{\sum_j W_j \tilde{X}_j^2 - (\sum_j W_j \tilde{X}_j) \tilde{X}_i}{\sum_j W_j \tilde{X}_j^2} \right)}{\sum_i W_i \left(\frac{(\sum_j W_j \tilde{X}_j^2) - (\sum_j W_j \tilde{X}_j) \tilde{X}_i}{\sum_j W_j \tilde{X}_j^2} \right)} \\ &= \frac{W_i \left(1 - \frac{(\sum_j W_j \tilde{X}_j) \tilde{X}_i}{\sum_j W_j \tilde{X}_j^2} \right)}{\sum_i W_i \left(1 - \frac{(\sum_j W_j \tilde{X}_j) \tilde{X}_i}{\sum_j W_j \tilde{X}_j^2} \right)} = \frac{\tilde{W}_i}{\sum_k \tilde{W}_k} \end{aligned}$$

where $\tilde{W}_i = W_i(1 - \frac{\sum_j W_j \tilde{X}_j}{\sum_j W_j \tilde{X}_j^2}\tilde{X}_i)$. Why is this tedious calculation useful? It tells us at least two useful things!

1. It implies that the weights sum to 1 (though they are not per se positive!).
2. It tells us that $\sum_i w_i \tilde{X}_i = 0$ (Exercise: show this)

More broadly, it is useful to have direct expressions for the weights because such expressions exposing what is really going on when we do different kinds of regression. Such weights can be plotted (see Problem Set) to give an intuitive visual picture as to which values of the running variable are being up-weighted or down-weighted. **Notice:** these weights are NOT AT ALL a function of the outcome variable Y. They are entirely based on the running variable and Kernel. They are applicable to any RD design you might run with the given Kernel and $\tilde{X}_j$ magnitudes.

6.4.2 Bias at the boundary

The expression $\hat{\alpha}_1 = \sum_{i=1}^{n_1} w_i Y_i$ is useful for considering the bias of the RD estimator.[16] The expected value of $\hat{\alpha}_1$ is then, assuming we have i.i.d. observations,

$$\mathbb{E}(\hat{\alpha}_1|X) = \sum_{i=1}^{n_1} w_i \mathbb{E}(Y_i|\tilde{X}_i) = \sum_{i=1}^{n_1} w_i \mu_1(\tilde{X}_i)$$

where we use the function $\mu_1(x)$ to denote the true functional form of $\mathbb{E}(Y_i(1)|\tilde{X}_i)$ (can sub in $Y(1)$ here because we are considering the treated group). We are interested in $\hat{\alpha}_1$ as an estimate of $\mathbb{E}(Y_i(1)|\tilde{X}_i = 0) = \mu_1(0)$, so the bias is

$$\text{bias} = \left(\sum_{i=1}^{n_1} w_i \mu_1(\tilde{X}_i)\right) - \mu_1(0)$$

To study this bias, let us do a Taylor Expansion of $\mu_1(\tilde{X}_i)$ about $\tilde{X}_i = 0$. We will assume that μ_1 is differentiable.

$$\mu_1(\tilde{X}_i) = \mu_1(0) + \mu_1'(0)\tilde{X}_i + \sum_{r=2}^{\infty} \frac{\mu_1^{(r)}(0)}{r!}\tilde{X}_i^k$$

When we substitute this into the expression for the bias, we get a nice cancellation of $\mu_1(0)$ terms in the overall expression (using the fact that the weights sum to 1) and the $\mu_1'(0)$ term (because $\sum_i w_i \tilde{X}_i = 0$

$$\begin{aligned}\text{bias} &= \mu_1(0) + \mu_1'(0)\sum_i w_i \tilde{X}_i + \sum_{r=2}^{\infty} \frac{\mu_1^{(r)}(0)}{r!}\sum_i w_i \tilde{X}_i^k - \mu_1(0) \\ &= \sum_{r=2}^{\infty} \frac{\mu_1^{(r)}(0)}{r!}\sum_i w_i \tilde{X}_i^r\end{aligned}$$

The implication of this cancellation is that bias is "zero up to first order" (the $\mu_1(0) - \mu_1(0)$ cancellation counts as 0^{th} order and always follows from doing the Taylor Expansion; the $\sum_i w_i X_i = 0$ is the first order part). If $\mu_1^{(k)}$ derivatives are bounded and the weights tend to be larger for $\tilde{X}_i$ close to 0 and small further away (as we'd hope) then we'd expect higher order terms to be increasingly small as we take $\tilde{X}_i$ close to 0 to higher powers. This is a big reason why people use local linear regression for regression discontintuity design rather than global polynomial regression!

In Module 6 Problem Set Question 3, you will show this result generalizes to unbiasedness up to order k for k^{th} order polynomial regression.[17] However, the fact that k^{th} order polynomials are unbiased up to k^{th} order does not justify making k as large as possible! We have to think about bias-variance trade-offs. High order polynomials have less bias but much more variance. You will show this in the problem set as well.

[16] Again, here I am re-indexing the treatment group to $i = 1, ..., n_1$. Everything can be applied to the $T_i = 0$ group, too.

[17] **Tip:** there is a way to derive the necessary properties of the weights w_i without the tedious element-by-element calculations of the previous section – keep things in matrix notation!

6.5 Problems and Diagnostic Tests

6.5.1 Problems - what can go wrong?

- **Competing interventions or shocks** also occuring at $X = c$ can create violations of continuity.

  Imagine that in a school, students who score $X \leq c$ on a pre-test at the start of the school year are both enrolled in tutoring and switched to a smaller classroom with more individualized support. Suppose we only know about the tutoring. We might then observe a big jump at $X = c$ that seems as if it reflects tutoring but actually reflects class assignment (e.g., if tutoring is really ineffective). This would look something like in Figure 11. We might call class size a 'competing intervention' which confounds our ability to study the effect of tutoring. Competing interventions which have no effect on outcome are not an issue, but it may be implausible to assume this. Other interventions which occur far from the cut-off (e.g., those who score below $d \ll c$ are sent to a different school) are not an issue for the continuity assumption, though we need to be careful that our estimators are not distorted by them (see bandwidth discussion).

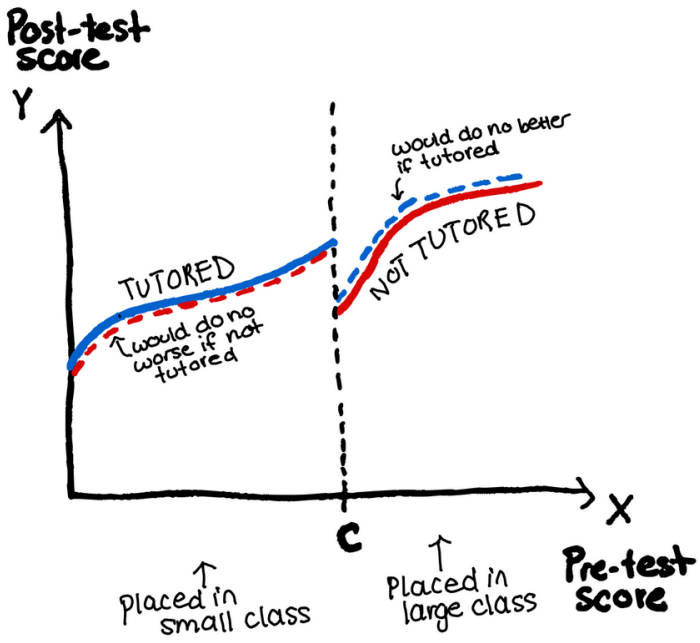
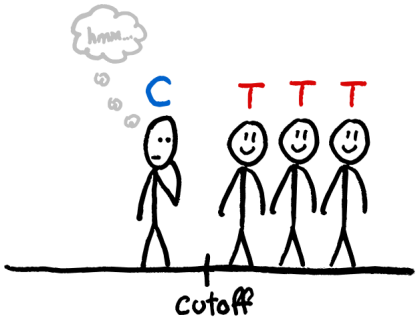


Figure 11: Illustration of violation of continuity assumption in a situation where treatment (tutoring) is totally ineffective on average but we would observe a difference in our data because the treat and control group vary by another variable (class size).

- **Sorting/Manipulation**: manipulation can happen when people know about the cut-off that determines treatment assignment, have an interest in being on one side or the other, and have the means to try to manipulate their X_i values to be on the desirable side. Our original deterministic sorting assumption implicitly says that people will be sorted according to their non-manipulated X_i values. If we instead have manipulated X_i values, this can lead to violations of continuity because $X_i^{manipulated}$ now represents something different and has a relationship to Y at the cut-off apart from the actual consequences of receiving/not receiving treatment.

  Common examples include people under/over reporting their income or age to be eligible for a benefit, prize, or scholarship. In the tutoring example, imagine that students (and their parents) know that if they receive a grade below c on the pre-test for a class, they will be eligible for free tutoring. Imagine that this leads some portion of students to purposefully do worse on the pre-test than they could in order to get tutoring and to aim for just below c so as not to make it *too* obvious they didn't give the test a fair attempt. We might then observed data where there are few observations just above c and a spike of observations just below c. Intuitively, the students just below the threshold (includes manipulators who would have scored high) might not be so similar to students who score just above (perhaps mainly students who actually got that score based on their best efforts). That lack of comparability ultimately threatens our continuity assumption.



6.5.2 Diagnostics

Diagnostic tests for RD are designed to help us check for each of the above problems.

1. **Competing Interventions - the Placebo Test**
   In this test, we plot the relationship between the running variable and other **pre-treatment** variables in our data and check whether there are any jumps at the cut-off $X = c$. For example, we might...

   - Check the relationship between the current X and a lagged outcome (Y in an earlier year when $X = c$ cut-off not yet applied).
   - Look at the relationship between X and Y in a different but comparable time or place where cut-off c was not in effect
   - Check whether there is any jump in covariates we believe might be related to outcome at the cut-off (e.g., age, race, education level).

   Observing a jump suggests that there is 'something else going on' that could lead to violations of the continuity assumption. In the example in Figure 11, if we had access to class size information and plotted it as a function of X, this would alert us to the other difference between the tutored and untutored groups. Overall, good placebos are ones that are thought to be related to the outcome of interest and are not expected to have any jump at $X = c$. Note that this test could also end up reflecting manipulation if, for example, the manipulators have certain common characteristics (e.g., wealthy individuals having more means to manipulate).
2. **Manipulation - McCrary Density Test** - this test looks for continuity in the distribution of the running variable around the cut-off. If some units are trying to get their X_i values above c (say), this can create a displacement that results in a bump in the distribution on one side of the threshold. This could be from people just under c pushing their value to $X_i > c$ (so that there is also a drop on the other side of c) or it could be people whose true/unmanipulated values are far from the cut-off (or both). See Figure 12 for what to look for.

For both these tests: getting a visual sense of what is going on is always wise, but to be more formal about what counts as a gap, we can also run a test in which we fit a local linear regression on either side of the threshold and look for a jump – i.e. again doing a regression discontinuity only now with a different variables (Placebo test) or with frequency counts as the outcome (Density test) and with the hope to *not* see a discontinuity.

Caution: for each of these diagnostic tests, we should think of them in line with hypothesis tests where failing to reject the null does not mean accepting that the null is true. If we run one of these diagnostic tests and observe a problem, we might conclude there is likely to be a violation of the RD assumptions. If we see no problem, this is a good sign, but it does not definitively tell us that there are no hidden violations. We may have looked at the wrong variables or simply not have access to relevant ones (e.g., in Figure 11, if we simply did not know anything about class size).

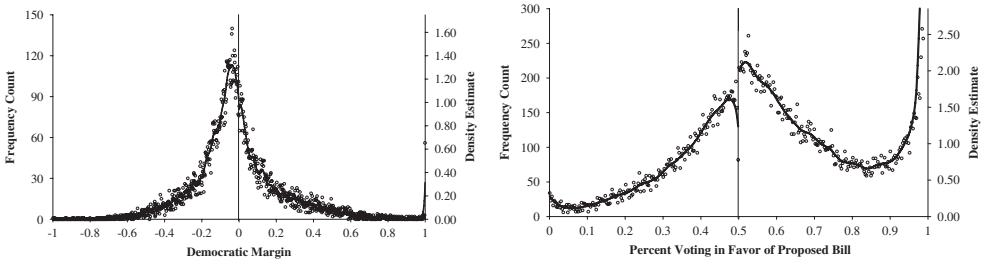


Figure 12: Examples for density test: taken from Kosuka Imai's lecture slides, Fall 2023. The right suggests manipulation.

6.6 Non-Assumptions

In RD, there are some things that might seem an issue but are allowed and some things that might seem required but are not:

- **Allowed - an relationship between X and Y**: Running variable X can have a causal or observational relationship to Y apart from treatment. For example, in Figure 9, we see Y increases with X for both groups, and in a different way for each group.
- **Not required - arbitrary rule**: the cut-off c need not be in any way arbitrary or selected 'by chance.' For example, SAT score cut-offs for college admissions, the alcohol drinking age, and the age required for Medicare health insurance in the U.S. are all real-world cut-offs with various justifications based at least in part on biology, social norms, political considerations etc. That is fine. That said, less arbitrary cut-offs may raise more concerns about manipulation because they may be widely known and socially important.
- **Not required - local randomization of X**: we do not require there to be any arbitrariness or randomness in who is on either side of the threshold c. A common intuition for RD is that around the threshold, there might be some 'as-if random' assignment to treatment, some element of chance in who ends up below c or above c so that we essentially have a mini experiment there. For example, for test scores, perhaps people just below c and above c only scored differently due to circumstantial factors on the day of the test. This *could* justify the RD approach by supporting continuity, but it is not necessary. Even if test scores are a perfect reflection of knowledge and ability, we might argue that students who score just below c and just above c are not likely to be so different that in the world whether neither were treated or both were treated, one group would have drastically higher potential outcomes than the other. That is: the continuity assumption says the *trends* in potential outcomes must be continuous, not that some people who are really of ability $> c$ ended up below c and vice versa.

  In fact, an estimation strategy based on as-if randomness would have disadvantages. Such a strategy might suggest taking a difference in means between the $X_i < c$ group and the $X_i \geq c$ group within some window where we think random assignment is plausible. This would implicitly assume the slope is 0 there. As shown in Figure 13, this could distort estimation if there is in fact a slope. In the figure, the left shows a bigger difference in window means while the right actually has a larger discontinuity!

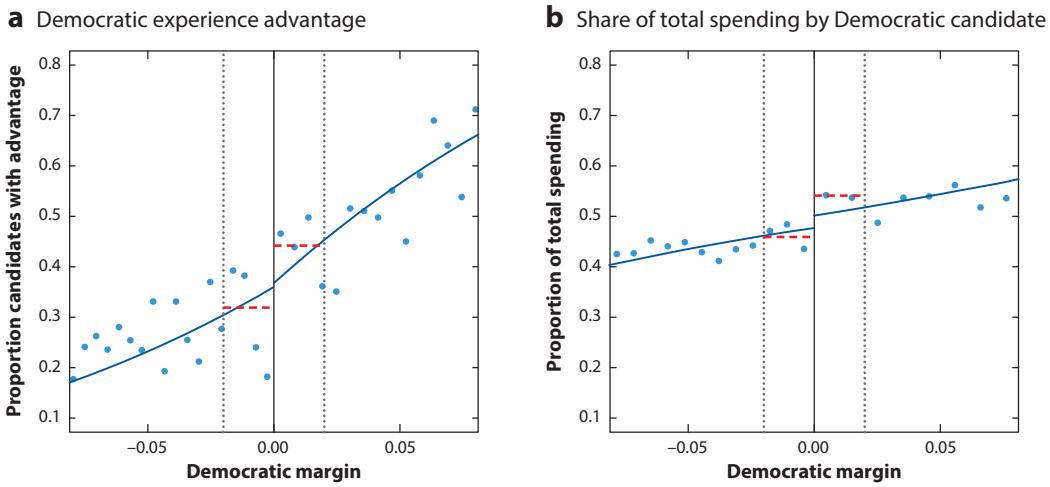


Figure 13: Figure from Kosuke Imai's Lecture Slides, Fall 2023

6.7 Limitation of RD: External Validity

Like instrumental variables, regression discontinuity brings us identification at a cost. A RD estimator really only identifies the effect of treatment for those with $X = c$. We might be tempted to argue that this effect extends at least near the threshold. We can always try to argue this, but it would be an assumption, not something we have learned from data. The further from the cut-off we go, the more untenable it may be to argue the causal effect at c is applicable to units with those values of X.

6.8 Further Resources

- Cunningham (2021) has a nice chapter on RDD with some real-world examples. You can also read more about Fuzzy RD designs there `https://mixtape.scunning.com/06-regression_discontinuity#the-sharp-rd-design`

6.9 Extension: Fuzzy RD

In a sharp RD design, we assume that the cut-off is deterministic. But in some real-world scenarios, it may be that actually, the cut-off is only an *encouragement* to get treatment as we discussed for instrumental variables in Module 5. For example, people with $X > c$ may become eligible for a benefit but some might not take the benefit. Or, people with $X > c$ might receive free tutoring but people with $X < c$ *could* pay and get tutoring, too. We can approach this scenario by combining the instrumental variables and RD perspectives.

Suppose $Z_i = I(X_i > c)$ is a deterministic encouragement ($Z_i = I(X_i \leq c)$ also possible). Then as in IV, we can think of there being principal treatment strata – always takers, never-takers, compliers, and defiers – defined by their $\{T(z) : z = 0, 1\}$ potential treatment outcome pair. We again have potential outcomes $Y_i(z, T_i(z))$. We make the following assumptions:

1. **Monotonicity** (from IV): $T_i(1) \geq T_i(0)$. This means that there are no people who only take treatment if not encouraged. Put another way, the only people on the non-treatment side of the cut-off who get treatment are always-takers.
2. **Exclusion** (from IV): $Y_i(1, t) = Y_i(0, t)$ for $t = 0, 1$ – if encouragement has no effect on treatment, it has no effect on outcome
3. **Continuity** (from RD): we now have to assume continuity in the expected potential outcome function both for treatment and outcome. That is, $\mathbb{E}(T_i(z)|X_i = x)$ is continuous in x for $z = 0, 1$ and $\mathbb{E}(Y_i(z, T_i(z))|X_i = x)$ is continuous in X for $z = 0, 1$. This assumption replaces randomization in the usual IV assumptions
4. **Relevance** (from IV): we assume the encouragement has some effect on treatment – i.e. compliers exist

Given these assumptions, it is possible to identify an average treatment effect for **compilers** at the **cut-off**.

$$\tau = \mathbb{E}(Y_i(1, T_i(1)) - Y_i(0, T_i(0)) \mid T_i(1) = 1, T_i(0) = 0, X_i = c)$$

The identification equality for this estimand is similar to the Wald Estimator: the ratio of the effect on Y and the effect on T, now each formulated as RD estimands via limits at the cut-off

$$\tau = \frac{\lim_{x \downarrow c} \mathbb{E}(Y_i|X_i = x) - \lim_{x \uparrow c} \mathbb{E}(Y_i|X_i = x)}{\lim_{x \downarrow c} \mathbb{E}(T_i|X_i = x) - \lim_{x \uparrow c} \mathbb{E}(T_i|X_i = x)}$$

A cost of moving to the fuzzy RD, IV type case is that the external validity picture grows even more limited. Now, we are not only estimating an effect only for those at the boundary but also only for compliers at the boundary, which could be an unusual and distinctive group.

To estimate the numerator and denominator, we again need to specify some functional form (e.g., use local linear regression) and then we fit the RD regressions for both Y_i and T_i and get an estimator of the form

$$\hat{\tau} = \frac{\hat{\alpha}_1 - \hat{\alpha}_0}{\hat{\gamma}_1 - \hat{\gamma}_0}$$

Where α_t are the intercepts for the Y_i regression and γ_t are the intercepts for the T_i regression.

Note that manipulation and competing interventions can still create violations of the above assumptions. Manipulation, for example, would mean units can select themselves to be encouraged or not.

6.10 Extension: RD in presence of manipulation

As we will see further in Module 7, if we cannot point identify an estimand, we can sometimes at least identify some bounds on what value it could take. Gerard et al. (2020) present methods for this in the context of RD design with manipulation of X_i values. Let $M_i = 1$ if unit i is a manipulator and $M_i = 0$ if not. If we add a **monotonicity assumption** that units only ever manipulate themselves to be on one side of the threshold, then we can identify the proportion of manipulators. This assumption can be realistic if there is one side of the cut-off that is clearly desirable for all units. For example, if people on side $X_i > c$ receive a scholarship, then there would be a clear incentive for some people to try to falsely boost their X_i values but it seems less likely that anyone would purposely lower their apparent X_i value to be below the cut-off.

Let $f(X_i = x)$ be the probability density of X_i. Under no-manipulation, we would expect this density to have no discontinuity at $X_i = c$. Let $f(X_i = c^-) = \lim_{x \uparrow c} f(X_i = x)$ be the limiting value approaching c from below and similarly define $f(X_i = c^+) = \lim_{x \downarrow c} f(X_i = x)$. Assume that if there is manipulation, it is monotonic as described above and of a type what will result in $f(X_i = c^-) < f(X_i = c^+)$.[18] Assume that in the absence of manipulation, the density would be continuous at c. Then intuitively, if there is a gap at c as in as in Figure 12, the size of that gap tells us about the proportion of manipulators. The proportion of manipulators is then identified as:

$$\lambda = P(M_i = 1 | X_i = c) = \frac{f(X_i = c^+) - f(X_i = c^-)}{f(X_i = c^+)}$$

or alternatively we could write

$$\lambda = P(M_i = 1 | X_i = c) = 1 - P(M_i = 0 | X_i = c) = 1 - \frac{f(X_i = c^-)}{f(X_i = c^+)}$$

To build intuition for why these work, imagine two tiny intervals width ϵ, one to the left of c and one to the right of c as in Figure 14. By monotonicity, the integral over the left interval represents the proportion of people in the population who are both non-manipulators and have true $X_i \in [c - \epsilon, c]$. The integral over the right interval represents the proportion of the population who are either manipulator or non-manipulator and have reported $X_i \in [c, c + \epsilon]$. Assuming continuity, as $\epsilon \to 0$, the left integral approaches the proportion of the population who are non-manipulators and report $X_i = c$ (truthfully) and the right integral approaches the proportion of the population who are manipulators or non-manipulators and report $X_i = c$. In the limit, we have:

$$\Pr(\text{non-manipulator} | X_i = c) = \frac{\Pr(\text{non-manipulator with } X_i = c)}{\Pr(X_i = c)} = \frac{\Pr(\text{non-manipulator with } X_i = c)}{\Pr(\text{non-manipulator or manipulator with } X_i = c)}$$

It is not strictly correct to write $\Pr(X_i = c)$ here because for a random variable with a density, technically $\Pr(X_i = x) = 0$ for any x, but this is a loose justification for the kind of thing we are getting at in the limit.

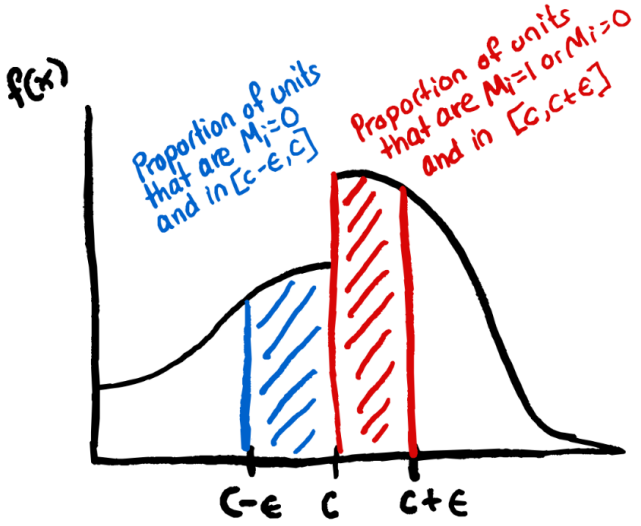


Figure 14: Illustration of intuition for identifying proportion of manipulators

[18] If manipulation is the other direction, we can flip things below. You can also imagine cases of manipulation where people who manipulate come from far below c and push their values far above c so that no discontinuity would arise at c itself. That kind of manipulation is not detected here but also may not be a problem.

Partial Identification:

Now consider the average treatment effect on non-manipulators as an estimand $\mathbb{E}(Y_i(1) - Y_i(0)|X_i = c, M_i = 0)$. Under monotonicity, for the no-treatment side of the cut-off,[19]

$$\mathbb{E}(Y_i(0)|X_i = c, M_i = 0) = \lim_{x \uparrow c} \mathbb{E}(Y_i|X_i < c)$$

However, for the treatment side, we have

$$\mathbb{E}(Y_i|X_i = c) = \mathbb{E}(Y_i(1)|X_i = c) = \lambda \mathbb{E}(Y_i(1)|X_i = c, M_i = 1) + (1 - \lambda)\mathbb{E}(Y_i(1)|X_i = c, M_i = 0)$$

and rearranging gives

$$\mu_1 := \mathbb{E}(Y_i(1)|X_i = c, M_i = 0) = \frac{\mathbb{E}(Y_i|X_i = c) - \lambda \mathbb{E}(Y_i(1)|X_i = c, M_i = 1)}{1 - \lambda}$$

λ and $\mathbb{E}(Y_i|X_i = c)$ are identifiable, but $\mathbb{E}(Y_i(1)|X_i = c, M_i = 1)$ is not. The partial identification approach is to consider the maximum and minimum values this could take. In this case, μ_1 is maximum if $\mathbb{E}(Y_i(1)|X_i = c, M_i = 1)$ is minimum, which happens if the $M_i = 1$ group represents the lowest λ quantile of $Y_i(1)$'s and μ_1 is minimum if the $M_i = 1$ group represents the highest λ quantile of $Y_i(1)$'s. Empirically, we can use this to estimate bounds on our quantity of interest by taking the average of the $n_1\lambda$ top and $n_1\lambda$ bottom treated observations and plugging each in for $\mathbb{E}(Y_i(1)|X_i = c, M_i = 1)$ above.

Manipulation in general:

Manipulation is a complex phenomenon and the above argument explicitly or implicitly assumes away some kinds of manipulation. What if, for example, we have a bunch of manipulators who try to get their X_i above c but actually only push it to just below c (a kind of fuzzy manipulation)? I do not believe the above analysis allows for that possibility. It might be that the assumption is not only that no one tries to manipulate themselves below the threshold but that no one just fails to manipulate themselves above the threshold. Perhaps it would be explore-able via other partial identification approaches! On the other hand, some kinds of manipulation are not a problem – at least for identification. If people far below the cut-off leapfrog themselves far above the cut-off, then these are not relevant to identifying the effect at $X_i = c$, though they could certainly affect estimation if they are within our bandwidth.

[19] Assuming here this is below c but that can be flipped.

7 Module 7: Regression with Observational Data

Big Picture: we arrive in the world of observational studies and ask a key question: when can we do a regression and recover causal effects? And what can go wrong? In particular, we consider settings where, unlike in regression discontinuity, people at any $X_i = x$ can receive any treatment (overlap) and explore the key notion of unconfoundedness, which, if met, can return us to something like the experimental setting. We also look at some options for if we do not entirely believe our assumptions: sensitivity analysis, modeling selection bias, and partial identification approaches.

Notation

- T_i treatment (assume binary for now, but ideas generalize)
- Y_i outcome
- X_i pre-treatment covariates

7.1 Identification Under (Un)confoundedness and Overlap Assumptions

Suppose we have observational data on (T_i, Y_i, X_i), meaning that the data did not come from a randomized experiment. For example, we might have school administrative records on children's math test scores, whether or not they received tutoring, and some other information about them such as their age, sex, grade, teacher, race, and socioeconomic status. Some key complications emerge here

- **Unknown treatment assignment mechanism**: we do not know why some people got treated and others did not
- **Confounding**: potential outcomes may no longer be independent of treatment, the result of a non-randomized assignment mechanism

$$T_i \not\perp\!\!\!\perp \{Y_i(1), Y_i(0)\}_{i=1}^n$$

In this setting, identification is only possible if we make further, additional assumptions. The two crucial ones are

1. **Overlap**

$$0 < \Pr(T_i = 1 | X_i = x) < 1 \quad \forall x$$

This says that units at every $X_i = x$ has some chance or receiving $T_i = 1$ or $T_i = 0$. Intuitively, for units that only ever receive $T_i = 1$ or only ever receive $T_i = 0$, we will never be able to make comparisons between treated and control units at that level of $X = x$. There is no element of random assignment there and any claims about causal effects at $X = x$ will be **extrapolations**.[20]

Checking this assumption can be hard. For a single discrete X_i, it may be doable. In the tutoring example, if for fifth-graders, X_i is teacher and there are 5 possible fifth grade teachers, it may be feasible to check whether, for each teacher, there are tutored and untutored students. If so, we conclude overlap. But if not, we technically have not proved overlap *does not hold*. Perhaps for one teacher, all children had some positive probability of getting tutoring, but by chance, none did. This all gets much harder when X_i is multivariate and includes continuous variables. The number of possible values of X_i quickly explodes and you would not expect to observe every possibility, let alone treatment and control present in every possibility. Should I be worried if, among the two female, Hispanic students of age 10 from fifth grade with teacher A, both got tutoring? Overall, it is possible to get *some sense* of how reasonable this assumption may be from the data but it may be hard or impossible to prove or disprove it. It helps here to have some knowledge of how treatment was assigned. For example, was tutoring advertised to everyone? Was it available at the school or in some location requiring extra transportation? Did teachers have to recommend students for it or could anyone enroll?

Adjusting estimand: if there is a group of units known have probability 0 or 1 of getting treatment, we could remove them from our data but this would change the interpretation of the estimand – the average treatment effect would not apply to them.

Problems with extremes: as we will see later when we talk about propensity scores, having probabilities very close to 0 or 1 can also be a problem for estimation.

[20] Recall: in RD designs, we actually accepted this some element of extrapolation via the continuity assumption. Overlap by construction does not hold for an RD design.

2. **Unconfoundedness**

$$T_i \perp \{Y_i(1), Y_i(0)\}_{i=1}^n \mid X_i = x \quad \forall x$$

- Also known as: exogeneity, ignorability, selection on observables, no omitted variable, and more
- Cannot be checked or tested because do not observe $Y_i(0), Y_i(1)$ together

This assumption is fundamental and tricky. The key idea is that conditional on some covariates, we essentially have a random experiment. It poses that we have access to *all* the factors X driving treatment assignment in systematic ways. In some contexts, this can be very reasonable. If we knew, for example, that a company's decision to grant a loan was based on five pieces of information about an applicant followed by a lottery, then conditioning on those five variables would achieve unconfoundedness. In reality, we often do not exactly know how treatment was assigned and have to try to evaluate whether we are conditioning on 'enough.'

> "Informally, unconfoundedness requires that we have a sufficiently rich set of pre-treatment variables so that adjusting for differences in values for observed pre-treatment variables removes systematic biases from comparisons between treated and control units. This critical assumption is not testable." (Imbens and Rubin, 2015, p.479)

One intuition for why we want unconfoundedness is that it provides us with **mini randomized experiments** within each $X_i = x$ group (this works if X is discrete - for continuous you could imagine a neighborhood).

Example: Confounding Imagine tutoring costs a lot of money so that only students from high SES backgrounds get tutoring, but these students tend to already have more advantages and hence have $Y_i(0) = Y_i(1) =$ pass math test while students without tutoring are some mix of $Y_i(0) = Y_i(1) = 1$, $Y_i(0) = 0, Y_i(1) = 1$ and $Y_i(0) = Y_i(1) = 0$ with a significant proportion who end up failing the test. Then we observe a huge apparent treatment effect driven in large part by students with no individual treatment effect! At the extreme, if all untutored students are $Y_i(0) = Y_i(1) = 0$, there'd be no true average effect. Notice here that the actual assignment mechanism (by SES) is not directly using $Y_i(0), Y_i(1)$. It cannot be because these are entirely unknown pre-treatment. However the assignment mechanism sorts people with different $Y_i(0), Y_i(1)$ into different groups – it creates correlation between $Y_i(0), Y_i(1)$ and T.

Example: Conditioning on X if, in the example above, SES is the only thing driving tutoring so that within a certain SES level, tutoring assignment is random and if, at each SES level, students still have *some chance* of getting tutoring, even if they are more likely to get it for higher SES levels, then we can get at causality by looking at the tutoring effect within SES levels.

Terminology Note: the pair unconfoundedness + overlap is sometimes referred to as **strong ignorability** while unconfoundedness alone is sometimes called **ignorability.**

Identification: Given unconfoundedness and overlap, the identification argument is a straightforward consequence of the Law of Total Expectation:

$$\begin{aligned}
\mathbb{E}(Y_i(1)) - \mathbb{E}(Y_i(0)) &= \mathbb{E}(\mathbb{E}(Y_i(1)|X_i)) - \mathbb{E}(\mathbb{E}(Y_i(0)|X_i)) \\
&= \mathbb{E}(\mathbb{E}(Y_i(1)|T_i = 1, X_i)) - \mathbb{E}(\mathbb{E}(Y_i(0)|T_i = 0, X_i)) \qquad \text{(unconfoundedness, overlap)} \\
&= \mathbb{E}(\mathbb{E}(Y_i|T_i = 1, X_i)) - \mathbb{E}(\mathbb{E}(Y_i|T_i = 0, X_i)) \qquad \text{(by consistency)} \\
&= \mathbb{E}(\mu_1(X_i) - \mu_0(X_i))
\end{aligned}$$

where $\mu_t(x)$ is the conditional expectation function from regressing outcome on treatment t and $X = x$. The last line above is entirely in terms of quantities estimable from observed data. That is, we can estimate both $\mu_1(x) = \mathbb{E}(Y_i|T_i = 1, X_i = x)$ and $\mu_0(x) = \mathbb{E}(Y_i|T_i = 0, X_i = x)$ from our data via regression and then average them using the empirical distribution of X_i's.

Note: previously, when we justified $\mathbb{E}(Y_i(1)|X_i = x) = \mathbb{E}(Y_i(1)|T_i = 1, X_i = x)$ by unconfoundedness, positivity was implicitly met simply because we had complete randomization. If $T_i = 1, X_i = x$ has probability 0, then the right side is not well defined while the left side is.

7.2 Regression Estimators

Following the above identification equality, two general **regression-based** estimators of the ATE are (see also Slide 4)

1. **Plug-in estimator** - uses only fitted values from the treatment and control regressions

$$\hat{\tau}_{reg} = \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) \tag{7}$$

2. **Imputation Estimator** - uses the original observed values $Y_i = Y_i(T_i)$ where possible and only imputes the missing potential outcome.

$$\hat{\tau}_{reg-imp} = \frac{1}{n}\sum_{i=1}^{n} T_i(Y_i - \hat{\mu}_0(X_i)) + (1-T_i)(\hat{\mu}_1(X_i) - Y_i) \tag{8}$$

Each of these estimators fits some regression $\hat{\mu}_t(X_i)$ to each of the $t = 0, 1$ groups and then takes the outer expectation over X_i relative to the empirical distribution of X.[21] These regressions can be linear models such as $Y_i(1) = \alpha + T_i\beta + X_i\gamma + \epsilon_i$ but also generalized linear models such as logistic regression in the case of a binary outcome or even the result of some machine learning technique. Of course unbiasedness of the estimator will depend on unbiasedness of the model in question. In later modules we will see a different approaches targeting the same estimand in different ways or combining regression approaches with others to form doubly-robust estimators (more later).

7.2.1 Equivalences in case of linear model

A nice property in the case where we pose a linear model is that we can skip calculating Equations 7 and 8 explicitly and instead use the coefficient from regression output. In particular $\hat{\tau}_{reg}$ is equivalent to the $\hat{\beta}$ from

$$Y_i = \alpha + T_i\beta + X_i\gamma + \epsilon$$

Similarly, under a linear model, $\hat{\tau}_{reg-imp}$ is equivalent to the coefficient of treatment $\hat{\beta}$ when we add an **interaction** between treatment and each (centered) co-variate $\tilde{X}_i = X_i - \bar{X}_i$

$$Y_i = \alpha + T_i\beta + \tilde{X}_i\gamma + \delta T_i\tilde{X}_i + \epsilon$$

In fact, we already saw this idea in Section 4.5.2 in the context of regression for experiments. As noted there, a nice aspect of the imputation estimator is that it reflects fitting regressions on treated and control units entirely separately so that we are not using the value of $Y_i = Y_i(1)$ for a treated unit to estimate $Y_i(0)$ and vice versa. In the linear model plug-in estimator case, this is not true.

Proof: *equivalences.* Given the linear model without interaction, the fitted values under treatment and control for the regression are $\hat{\mu}_1(X_i) = \hat{\alpha} + \hat{\beta} + X_i\hat{\gamma}$ and $\hat{\mu}_0(X_i) = \hat{\alpha} + X_i\hat{\gamma}$. Taking the difference leaves only the β so we get $\hat{\tau}_{reg} = \frac{1}{n}\sum_{i=1}^{n}\hat{\beta} = \hat{\beta}$. To see the imputation estimator result, let $r_{i1} = Y_i(1) - \hat{Y}_i(1)$ and $r_{i0} = Y_i(0) - \hat{Y}_i(0)$. We then have

$$\begin{aligned}
Y_i(1) - \hat{Y}_i(0) = Y_i(1) - \hat{Y}_i(1) + \hat{Y}_i(1) - \hat{Y}_i(0) &= r_{i1} + \hat{Y}_i(1) - \hat{Y}_i(0) = r_{i1} + (\hat{\alpha} + \hat{\beta} + \tilde{X}_i\hat{\gamma} + \tilde{X}_i\hat{\delta}) - (\hat{\alpha} + \tilde{X}_i\hat{\gamma}) \\
&= r_{i1} + \hat{\beta} + \tilde{X}_i\hat{\delta} \\
\hat{Y}_i(1) - Y_i(0) = \hat{Y}_i(1) - \hat{Y}_i(0) + \hat{Y}_i(0) - Y_i(0) &= \hat{\beta} + \tilde{X}_i\hat{\delta} - r_{i0}
\end{aligned}$$

Plugging these into the regression-imputation estimator, we get

$$\begin{aligned}
\hat{\tau}_{reg-imp} &= \frac{1}{n}\sum_{i=1}^{n} T_i(\hat{\beta} + \tilde{X}_i\hat{\delta} + r_{i1}) + (1-T_i)(\hat{\beta} + \tilde{X}_i\hat{\delta} - r_{i0}) \\
&= \frac{1}{n}\hat{\beta}\sum_{i=1}^{n}(T_i + 1 - T_i) + \frac{1}{n}\sum_{i=1}^{n}\tilde{X}_i\hat{\delta}(T_i + 1 - T_i) + \frac{1}{n}\sum_{i=1}^{n}T_i r_{i1} + \frac{1}{n}\sum_{i=1}^{n}(1-T_i)r_{i0} \\
&= \hat{\beta} + \left(\frac{1}{n}\sum_{i=1}^{n}\tilde{X}_i\right)\hat{\delta} + 0 + 0 = \hat{\beta}
\end{aligned}$$

Where $\sum_{i=1}^{n}\tilde{X}_i = 0$ because of centering. The residuals sum to 0 because, letting $r_i = Y_i - \hat{Y}_i$ (residual on the observed data), we have $\sum_{i=1}^{n}T_i r_{i1} + \sum_{i=1}^{n}(1-T_i)r_{i0} = \sum_{i=1}^{n} r_i$ and it is a standard regression result that the residuals always sum to 0. □

[21] The empirical distribution is just a discrete distribution with weight $\frac{1}{n}$ for each X_i. It is an approximation of the (possibly continuous) distribution of X in the population and converges to that distribution as $n \to \infty$. Here we use it to approximate the outer expectation in the identification equality.

7.2.2 Other Estimands

Sometimes, instead of the ATE, researchers are interested in the **ATC** (average treatment effect on the controls) or **ATT** (average treatment effect on the treated). Theoretically, these are:

$$\mathbb{E}(Y_i(1) - Y_i(0)|T_i = 1)$$
$$\mathbb{E}(Y_i(1) - Y_i(0)|T_i = 0)$$

When we were studying experiments, these estimands were theoretically identically because we were randomly assigning to treatment and control. Now, with the introduction of X_i, it may be that the treatment and control group differ in their X_i profile (e.g., more older people get treated) so that even if unconfoundedness holds (randomizatin within X_i group), it might be meaningful to consider the groups separately by averaging over their different distributions of X_i. The identification equality for the ATT becomes (similarly for ATC):

$$\begin{aligned}
\mathbb{E}(Y_i(1)) - \mathbb{E}(Y_i(0)|T_i = 1) &= \mathbb{E}(Y_i(1)|T_i = 1) - \mathbb{E}(\mathbb{E}(Y_i(0)|T_i = 1, X_i)|T_i = 1) \\
&= \mathbb{E}(Y_i(1)|T_i = 1) - \mathbb{E}(\mathbb{E}(Y_i(0)|T_i = 0, X_i)|T_i = 1) \qquad \text{(unconfoundedness)} \\
&= \mathbb{E}(Y_i|T_i = 1) - \mathbb{E}(\mathbb{E}(Y_i|T_i = 0, X_i)|T_i = 1) \qquad \text{(by consistency)}
\end{aligned}$$

We again have imputation and plug-in estimators (here for ATT, similarly for ATC):

$$\hat{\tau}_{reg,ATT} = \frac{1}{n_1}\sum_{i=1}^{n} T_i(\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)$$

$$\hat{\tau}_{regimp,ATT} = \frac{1}{n_1}\sum_{i=1}^{n} T_i(Y_i - \hat{\mu}_0(X_i))$$

Note that in the case of the ATT (and ATC), we condition on the treatment (control) groups so in calculations, we generally consider the number of treated units n_1 (and hence n_0), as fixed even if it is not something pre-set by an experimenter (which was the logic for considering n_1 as fixed in Modules 1-4). See for example Module 8 Review Question 1.

7.2.3 Variance Estimation

1. **Analytic Calculation for τ_{reg}**
   The **conditional** variance of the regression estimator is given by the fairly complicated expansion

$$\begin{aligned}
\mathbb{V}(\hat{\tau}_{reg}|X) &= \frac{1}{n^2}\mathbb{V}(\sum_{i=1}^{n} \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)|X) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{V}(\hat{\mu}_1(X_i)) + \mathbb{V}(\hat{\mu}_0(X_i)|X)) + 2\text{Cov}(\hat{\mu}_1(X_i)\hat{\mu}_0(X_i)|X) + \sum_{i\neq j} \text{Cov}(\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i), \hat{\mu}_1(X_j) - \hat{\mu}_0(X_j)|X)
\end{aligned}$$

   This expression highlights that there is a lot of dependency coming from the fact that the regression parameters used to create each fitted value come from the same data. The first covariance reflects the idea that for a given unit, predictions for treatment and control may be correlated and the second covariance reflects correlation across observations in their fitted values. Especially for complex models for $\hat{\mu}_t$ (e.g., some non-linear function of covariates), calculating these quantities could be very tedious.

2. **Bootstrap** - instead of wrestling with the above expressions, it is common to bootstrap by re-sampling the n observations with replacement some B times, calculating $\hat{\tau}_{reg}$ (or $\hat{\tau}_{reg-imp}$) on each, and looking at the variance of the $\hat{\tau}_{reg}^{(k)}$ for $k = 1, ..., K$. Note that this gives an estimate of the **unconditional variance** (not conditioning on X). A drawback of this approach is that especially for complex models, it can be quite computationally intensive.

3. **Other options**: See, for example, Slide 5 for a reference on Quasi-Bayesian Monte Carlo methods that rely on a normal approximation argument to re-sample parameters $\hat{\alpha}, \hat{\beta}, \hat{\gamma}$ from a normal distribution and then estimate $\hat{\tau}_{reg}^{(k)}$ (or $\hat{\tau}_{reg-imp}$) for each parameter sample. This will still require some analytic computations to figure out the asymptotic variance of these parameters (e.g., using the delta method) but could be less difficult than dealing with the quantity above.

7.3 Sensitivity Analysis

The approach above should seem a bit suspect. Just assume unconfoundedness and positivity and then we get causality from correlation after all? Is that EVER realistic? And we can't even check unconfoundedness and positivity from our data! Maybe we can use our domain knowledge to reason they are likely to hold at least approximately? But then, how approximately is enough? Sensitivity analysis shifts the conversation away from a binary "Do the assumptions hold?" question and to the questions like "How sensitive are my conclusions to violations of the assumptions?" or "If there were truly no causal effect, how big would this confounding dynamic have to be to fully account for the observed effect?" It is an approach that deals in hypotheticals to try to better understand what kinds of things could be or are unlikely to be going on.

There are many approaches to sensitivity analysis. In this course, we briefly look at:

1. **Risk-Based Approach**: the classical approach based on decomposing risk ratios involving an omitted variable
2. **Regression-based Approach**: examines role of of omitted variables via omitted variable results for regression

7.3.1 Risk-Based Approach

The canonical example for this approach, which comes from Cornfield et al. (1959), is that of smoking and lung cancer. We cannot ethically do an experiment to establish whether smoking causes lung cancer. Given observational data only, it could be that there are other factors (e.g., genetics) which lead people both to smoke and to get lung cancer without the smoking actually causing cancer. The sensitivity analysis question is: how influential would those other factors have to be to *entirely* explain the observed smoking-cancer correlation? In this case, Cornfield et al. (1959) concluded that the effect of other factors would have to be implausibly large. Let T be an indicator for smoking, Y be a binary indicator for having lung cancer, and U be an unobserved confounder, which we will also encode as 0 or 1 (present or not present).[22] The Cornfield approach involves the following ratios:

1. **Actual risk ratio** $RR_{TY} = \frac{\Pr(Y_i(1)=1)}{\Pr(Y_i(0)=1)}$ – risk of getting cancer if smoke vs getting cancer if do not smoke
2. **Observed risk ratio** $RR^{Obs}_{TY} = \frac{\Pr(Y_i=1|T_i=1)}{\Pr(Y_i=1|T_i=0)}$ – ratio of observed proportion with cancer among smokers and non-smokers
3. **Risk of Y given U** $RR_{UY} = \frac{\Pr(Y_i=1|U_i=1)}{\Pr(Y_i=1|U_i=0)}$ – risk of cancer given hidden factor present or not
4. **Risk of U given T** $RR_{TU} = \frac{\Pr(U_i=1|T_i=1)}{\Pr(U_i=1|T_i=0)}$ – risk of hidden factor present given smoking or not

We then make the following assumptions:

1. **Unconfoundedness** given the hidden factor $Y_i(0), Y_i(1) \perp T|U$ (could also condition on observed covariates X here). This says treatment assignment is at random, at least among those with same U. It does not make any claims about T having or not having an effect on Y.
2. **Cornfield Condition**:

$$\frac{P(Y_i(1)=1|U=u)}{P(Y_i(0)=1|U=u)} = 1$$

   This says that conditional on people with $U=u$, there is no effect of smoking on cancer – i.e. the probability of getting cancer is the same regardless of smoking.

The question is: under these assumptions, how large would RR_{UY} and RR_{UT} have to be produce the observed RR^{Obs}_{TY} given that Cornfield Condition says actual conditional risk ratio is 1? The answer turns out to be:

$$\min(RR_{TU}, RR_{UY}) \geq RR^{obs}_{TY} \tag{9}$$

That is, both ratios need to be at least as big as the observed one! For example, if in the smoking context, the observed risk ratio is 10 (smokers 10 times more likely to have cancer), then we would need to have a U that makes people with $U=1$ at least 10 times more likely to get cancer and for which smokers ($T=1$) are 10 times more likely to have $U=1$ than non-smokers. [23] To give some idea of how Equation 9 can be derived, first note that the two assumptions above together imply that observationally,

$$Y \perp T|U$$

because

$$\begin{aligned}
P(Y=1|T=1,U=u) &= P(Y(1)=1|T=1,U=u) \quad \text{(consistency)} \\
&= P(Y(1)=1|U=u) \quad \text{(unconfounded)} \\
&= P(Y(0)=1|U=u) \quad \text{(cornfield)} \\
&= P(Y(0)=1|T=0,U=u) \quad \text{(unconfounded)} \\
&= P(Y=1|T=0,U=u) \quad \text{(consistency)}
\end{aligned}$$

[22] Note: there are extensions to the non-binary case. See Ding and Vanderweele (2014).
[23] In fact, an even tighter lower bound can be derived. See Ding and Vanderweele (2014).

Using this, we have:

$$
\begin{aligned}
RR_{TY}^{obs} &= \frac{P(Y=1|T=1)}{P(Y=1|T=0)} \\
&= \frac{P(Y=1|U=1,T=1)P(U=1|T=1) + P(Y=1|U=0,T=1)P(U=0|T=1)}{P(Y=1|U=1,T=0)P(U=1|T=0) + P(Y=1|U=0,T=0)P(U=0|T=0)} && \text{(Law of Total Probability)} \\
&= \frac{P(Y=1|U=1)RR_{TU} + P(Y=1|U=0)\frac{1-P(U=1|T=1)}{P(U=1|T=0)}}{P(Y=1|U=1) + P(Y=1|U=0)\frac{1-P(U=1|T=0)}{P(U=1|T=0)}} && \text{(apply } Y \perp T|U\text{, divide num. and denom. by } \Pr(U=1|T=0)) \\
&= \frac{P(Y=1|U=1)RR_{TU} + P(Y=1|U=0)\left(\frac{1}{P(U=1|T=0)} - RR_{TU}\right)}{P(Y=1|U=1) + P(Y=1|U=0)\left(\frac{1}{P(U=1|T=0)} - 1\right)} && \text{(simplify)}
\end{aligned}
$$

The last line can ultimately be re-arranged to get the equality:

$$RR_{TU} = RR_{TY}^{obs} + (RR_{TY}^{obs} - 1)\frac{P(Y=1|U=0)}{P(U=1|T=0)}$$

WLOG, let $RR_{TY}^{obs} \geq 1$.[24] Then the equality says that $RR_{TU} \geq RR_{TY}^{obs}$. This shows one part of equation 9. Crucially, this was derived under a no-effect-of-smoking condition so we are saying here that IF there is no effect of smoking, then RR_{TU} MUST be at least as large as than the observed risk ratio RR_{TY}^{obs} (where the riskier condition is in the numerator).

[24] If < 1, we can flip all the ratios to treat non-smoking as the higher-risk condition and get same results in terms of explaining the observed higher risk of non-smoking

7.3.2 Regression-Based

In the risk-based approach above, we posed a binary omitted variable U and used it in risk ratios. Another sensitivity analysis approach is to examine what happens if the true linear model is

$$Y_i = \alpha + \beta T_i + \gamma^T X_i + \delta U_i + \epsilon_i$$

for unobserved confounder U_i. We will assume here that U_i is a scalar.

Omitted Variable Bias: recall that in linear regression, if we fit the model $Y_i = \alpha + \beta T_i + \gamma^T X_i + \tilde{\epsilon}_i$ where $\tilde{\epsilon}_i = \delta U_i + \epsilon_i$ and if U_i and X_i are correlated, then the exogeneity assumption that $\mathbb{E}(\tilde{\epsilon}_i|X_i) = 0$ does not hold.[25] As a result, there is bias and one can show that the bias has the following expression:

$$\text{bias} = \mathbb{E}(\hat{\beta}|X,T,U) - \beta = \delta * \frac{\text{Cov}(T_i^{\perp X}, U_i^{\perp X})}{\mathbb{V}(T_i^{\perp X})}$$

In theory (if we had U), we can obtain a consistent estimate this bias by plugging in $\hat{\delta}$ from the full regression and the sample covariances and variances:

$$\widehat{\text{bias}} = \mathbb{E}(\hat{\beta}|X,T,U) - \beta = \hat{\delta} * \frac{\widehat{\text{Cov}}(T_i^{\perp X}, U_i^{\perp X})}{\hat{\mathbb{V}}(T_i^{\perp X})}$$

If, conditional on X and T, there is no association between the outcome and U_i so that $\delta = 0$ or if $T_i \perp U_i|X$ so that covariance above is 0 (we get this if unconfoundedness as in Section 7.1 holds), then there is no bias.

$A^{\perp B}$ Notation: Above, $A^{\perp B}$ is a shorthand for the residuals from regressing variable A on variable(s) B. Cov and $\mathbb{V}$ denote the covariance and variance of these residuals and the hat versions represent the sample covariance and variance. That is, $\hat{\mathbb{V}}(T_i^{\perp X}) = \frac{1}{n-1}\sum_{i=1}^n (T_i - \hat{T}_i)^2$ and $\widehat{Cov}(T_i^{\perp X}, U_i^{\perp X}) = \frac{1}{n-1}\sum_{i=1}^n (T_i - \hat{T}_i)(U_i - \hat{U}_i)$ where $\hat{T}_i$ comes from regressing T on X and $\hat{U}_i$ from regressing U on X.[26] This is the notation used in Cinelli and Hazlett (2019) except they denote the sample quantities without the hat notation – I add the hat to make the distinction clearer but note that the $\mathbb{V}$ without a hat may refer to the empirical quantity on the slides or review questions. See Appendix B for derivation of the omitted variable bias formula and for information on the Frisch-Waugh-Lovell (FWL) Theorem, which is the underlying reason for regressing on X first here.

Partial R^2: Recall that in for a standard regression of Y on X, the expression for R^2 is

$$R^2 = 1 - \frac{\text{Residual Sum of Squares (RSS)}}{\text{Total Sum of Squares (TSS)}}$$

and in the case of univariate X, Y, we have

$$R^2 = Corr^2(X,Y)$$

where this holds for theoretically and empirically (see Appendix B for derivation). Partial R^2 captures the same notion only in a case where we have already regressed on some variable Z and, in line with the FWL Theorem, are considering the remaining variation in Y explained by X after regressing both on Z. We denote this $R^2_{Y\sim X|Z}$ and there are a few different ways to write this expression. First, we have

$$R^2_{Y\sim X|Z} = \frac{R^2_{Y\sim X+Z} - R^2_{Y\sim Z}}{1 - R^2_{Y\sim Z}}$$

The numerator says that the amount of additional variation explained by X is the 'amount of variation explained by X and Z' minus the 'amount explained by Z.' The denominator is the leftover variation in Y *not* explained by Z.[27] The subscript for each R^2 term represents the OLS regression of the variable left of the $\sim$ on the variables right of the $\sim$. Let RSS stand for residual sum of squares. Then another way of writing the (empirical) partial R^2 is:

$$R^2_{Y\sim X|Z} = \frac{1 - \frac{RSS_{Y\sim X+Z}}{TSS} - 1 + \frac{RSS_{Y\sim Z}}{TSS}}{\frac{RSS_{Y\sim Z}}{TSS}} = 1 - \frac{RSS_{Y\sim X+Z}}{RSS_{Y\sim Z}} = 1 - \frac{\hat{\mathbb{V}}(Y^{\perp X,Z})}{\hat{\mathbb{V}}(Y^{\perp Z})}$$

Note also that in the case where X is univariate, we have

$$R^2_{Y\sim X|Z} = R^2_{X\sim Y|Z} = Corr^2(Y^{\perp Z}, X^{\perp Z})$$

a generalization of the aforementioned equivalence.

[25] See Appendix B.

[26] Note: this sum of squared residuals is exactly the empirical variance of the residuals because the residuals in a regression always have empirical mean 0.

[27] Reminder: "explained" here is not at all causal. This is just a standard way of describing how much one variable linearly varies with (linearly predicts) the other.

Partial R^2 Formulation of Bias Magnitude: as shown in Cinelli and Hazlett (2019), the estimated bias expression above can be re-written in way that highlights what it depends on.

$$(\widehat{\text{bias}})^2 = R^2_{Y\sim U|T,X} * \frac{R^2_{T\sim U|X}}{1 - R^2_{T\sim U|X}} * \frac{\hat{\mathbb{V}}(Y_i^{\perp X,T})}{\hat{\mathbb{V}}(T_i^{\perp X})} \tag{10}$$

where

- $R^2_{Y\sim U|T,X}$ has to do with how much additional variation in Y the unobserved U explains conditional on T, X.

  $$R^2_{Y\sim U|T,X} = \frac{R^2_{Y\sim U+T+X} - R^2_{Y\sim T+X}}{1 - R^2_{Y\sim T+X}}$$

  In theory, if we knew U, we could plug in these residual sum of squares to calculate the partial R^2. However, this quantity is actually **unobservable**. Notice that if U explains all the rest of the variation in Y, we get a 1 and if U explains no additional variation, we get 0.

- $\frac{R^2_{T\sim U|X}}{1-R^2_{T\sim U|X}}$ captures the relationship between U and T, X. If, conditional on X, U explains no variation in T, then the numerator above is 0 and the omitted variable bias is 0. This again reflects the idea that omitted variables only matter if they are correlated with the included covariates. We again have $R^2_{T\sim U|X} = 1 - \frac{RSS_{T\sim U+X}}{RSS_{T\sim X}}$. This term relates to the strength of confounder U. If we have some unobserved factor which is correlated with treatment (this term) and with outcome (previous term), we have a problem. Because this term involves U it is also **unobservable**.

- $\frac{\hat{\mathbb{V}}(Y_i^{\perp X,T})}{\hat{\mathbb{V}}(T_i^{\perp X})}$ is a ratio of residual variances.

  - The numerator is $\frac{1}{n-1}\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$ where $\hat{Y}_i$ comes from regressing Y on X, T. The bias is smaller if this residual is smaller intuitively because in this case there is not much 'room' left for a biasing confounder. In the extreme, if we have a really rich set of covariates X and T so that they perfectly predict outcome, then there can be no U_i with $\delta \neq 0$.
  - The denominator is $\frac{1}{n-1}\sum_{i=1}^n (T_i - \hat{T}_i)^2$ where $\hat{T}_i$ comes from regressing T on X. The bias is smaller if this is *larger* because this reflects more exogeneity in T_i. Think of our randomized experiment ideal: randomness in treatment assignment is a good thing! If treatment were some deterministic function of X, this would violate our positivity assumption.

  Overall, this quantity is something we can calculate from our data since it does not involve U/

Proof Sketch (for full version, see Module 7 Review Sheet 1)

- By FWL theorem, $\hat{\delta}$ from regression of Y on X, T, U is equivalent to regressing Y and U on T, X first and then regressing residuals on each other. By the covariance formulation of regression coefficients, we then have:

  $$|\hat{\delta}| = \frac{|\widehat{\text{Cov}}(Y_i^{\perp T,X}, U_i^{\perp T,X})|}{\hat{\mathbb{V}}(U_i^{\perp T,X})} = \frac{|\widehat{\text{Cov}}(Y_i^{\perp T,X}, U_i^{\perp T,X})|}{\sqrt{\hat{\mathbb{V}}(U_i^{\perp T,X})}\sqrt{\hat{\mathbb{V}}(Y_i^{\perp T,X})}} * \frac{\sqrt{\hat{\mathbb{V}}(Y_i^{\perp T,X})}}{\sqrt{\hat{\mathbb{V}}(U_i^{\perp T,X})}} = \sqrt{R^2_{Y\sim U|T,X} * \frac{\hat{\mathbb{V}}(Y_i^{\perp T,X})}{\hat{\mathbb{V}}(U_i^{\perp T,X})}}$$

  Where the last equality comes from an equivalent expression for partial R^2 in terms of squared correlation. That is, the standard result that for univariate regression of Y on X, $R^2 = Corr^2(X, Y)$ (theoretically and empirically) extends to partial R^2.

- Similarly, taking the second part of the estimated bias expression

  $$\frac{|\widehat{\text{Cov}}(T_i^{\perp X}, U_i^{\perp X})|}{\hat{\mathbb{V}}(T_i^{\perp X})} = ... = \sqrt{R^2_{T\sim U|X} * \frac{\hat{\mathbb{V}}(U_i^{\perp X}}{\hat{\mathbb{V}}(T_i^{\perp X})}}$$

- Multiplying these and squaring both sides gives the result after some manipulations.

How is this useful for sensitivity analysis?

We can use it to explore hypothetical possibilities! Although only the third term above is calculable from data, we can pose hypothetical values of the partial R^2 in the first and second term and examine how the magnitude of the bias would change over a range of values. The results are often plotted in a contour plot as in the plot on Slide 7, reproduced below. This plot does not tell us how big the bias actually is, but it does tell us **how big $R^2_{Y\sim U|T,X}$ and $R^2_{T\sim U|X}$ would have to be for the bias to be large enough to completely remove our observed effect.** Sometimes, people shade the plot to indicate the region in which bias would be large enough to flip the sign of the effect and use the area of this region as an indication of sensitivity.

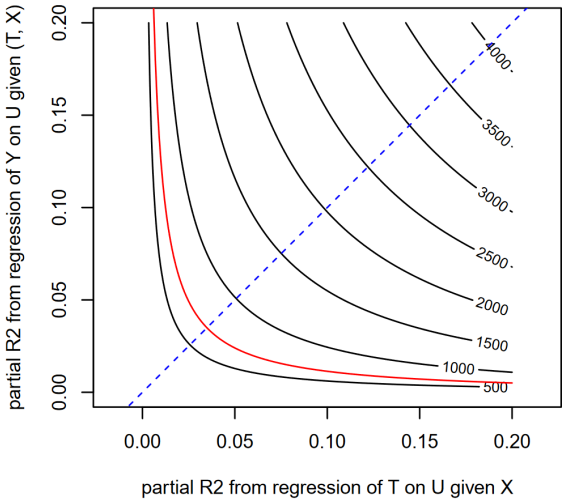


Figure 15: Example sensitivity analysis plot. In the context of the study this plot came from, there was experimental data that could be used to evaluate the 'true bias'. The curve represents compatible values for the two observable quantities in Equation 10 once we fix that bias.

7.4 Modelling Selection Bias

The sensitivity analysis in the previous section did not tell us how big the resulting bias actually is or allow us to try to get rid of it – it is more of a hypothetical exploration. Other work has tried to adjust for unobserved confounders, but this is challenging and requires its own assumtpions. We can think of the issue of confoundedness as a kind of **selection bias** in which, instead of people being randomly sorted into treatment and control, there are some factors pushing certain people towards treatment or control. An idea for making progress here is that there could be some unobserved U_i which has some dependence with observed X. We might then find some function of X which, if controlled for (aka incorporated into the regression), allows us to meet the unconfoundedness (exogeneity) condition.

As a simple illustration, imagine that we have Y, X and T, but actually, there is an unobserved confounder $U_i = X_i^2 + \epsilon$ where ϵ is some random noise. Then if we instead regress Y on X, X^2, T, we might get closer to unconfoundedness! In this case, although we were already using X to explain variation in Y, we were not fully leveraging it to create our 'mini' randomized experiments within units with similar covariates. It's as if there is **something we forgot to control for** but that something is encoded in our observed covariates.

7.4.1 Classic Approach: Heckman Model

There is a lot of work in economics on modelling selection bias and the canonical model is one from Heckman (1979). The set-up is as follows:

- **Outcome model:** $Y_i = \alpha + \beta T_i + \gamma^T X_i + \epsilon_i$ using just observed variables
- **Selection Model**: $T_i = 1\{T_i^* > 0\}$ with $T_i^* = \lambda + X_i^T\delta + \eta_i$. We will let $\eta_i | X_i \overset{iid}{\sim} N(0,1)$ to make this a **probit** model
  - What this says is that we are modeling a binary outcome (treatment or not) as a function of some continuous latent variable T^* which is a linear function of our observed X_i's and some (unobserved) η_i. In particular,

$$\begin{array}{rcccl} T_i = 1 & \leftrightarrow & T_i^* > 0 & \leftrightarrow & \eta_i > -\lambda - X_i^T\delta \\ T_i = 0 & \leftrightarrow & T_i^* \leq 0 & \leftrightarrow & \eta_i \leq -\lambda - X_i^T\delta \end{array}$$

  - The **probit model** is a generalized linear model for a binary outcome where, letting Φ be the cdf of a standard normal,

$$\begin{aligned} P(T_i = 1|X) &= P(T_i^* > 0|X) = P(\eta_i > -\lambda - X_i^T\delta|X) = \Phi(\lambda + X_i^T\delta) \\ P(T_i = 0|X) &= 1 - P(T_i = 1|X) = \Phi(-\lambda - X_i^T\delta) \end{aligned}$$

    Note that $P(T_i = 1|X)$ is not linear in X though T_i^* is.

Selection Bias: is reflected in the model above if $\mathbb{E}(\epsilon_i|T_i, X_i) \neq 0$ (a violation of exogeneity assumption). This could happen if ϵ_i and η_i above are correlated, even conditional on X. For example, imagine there is an unobserved U_i that plays a role in both the ϵ_i and η_i. Specifically, imagine that even among units with a fixed $X_i = x$, U_i leads to high η_i for certain units, which means a higher probability of $T_i = 1$. Imagine, at the same time, U_i leads to higher Y_i values for $X_i = x$. Then ϵ_i will also tend to be positive and large. Overall, we could have $Cov(\epsilon_i, \eta_i|X_i = x) > 0$ and $\mathbb{E}(\epsilon_i|T_i = 1, X_i = x) > 0$. This is essentially the usual intuition that it would be a problem to have an unobserved confounder affecting both treatment and outcome, but now we've represented it in terms of a specific model. Formally, consider the following expectations given the models above:

$$\begin{aligned} \mathbb{E}(Y_i|X_i, T_i = 1) &= \alpha + \beta + \gamma^T X_i + \mathbb{E}(\epsilon_i|X_i, T_i = 1) &&= \alpha + \beta + \gamma^T X_i + \mathbb{E}(\epsilon_i|X_i, \eta_i > -\lambda - \delta^t X_i) \\ \mathbb{E}(Y_i|X_i, T_i = 0) &= \alpha + \gamma^T X_i + \mathbb{E}(\epsilon_i|X_i, T_i = 0) &&= \alpha + \gamma^T X_i + \mathbb{E}(\epsilon_i|X_i, \eta_i \leq -\lambda - \delta^t X_i) \end{aligned}$$

Notice that the expectations above are always conditional on X_i, so we can view them as a function $g(X_i, T_i = 1) = \mathbb{E}(\epsilon_i|X_i, \eta_i > -\lambda - \delta^T X_i)$ and similarly for $g(X_i, T_i = 0)$. We could then imagine the regression

$$Y_i = \alpha + \beta T_i + \gamma^T X_i + \kappa g(X_i, T_i) + \xi_i \tag{11}$$

where ξ_i truly is some exogeneous noise and $g_{T_i}(X_i)$ is essentially that 'something' we forgot to control for. In this sense, the selection bias can be viewed as a model **specification error**. We do not actually know g because we do not know ϵ_i values or their expectation (nor η or δ), but could we somehow estimate it?

Heckman's approach is to derive this term under a **bivariate normal assumption**:

$$\begin{bmatrix} \epsilon_i \\ \eta_i \end{bmatrix} | X_i \overset{iid}{\sim} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{bmatrix} \right)$$

here ρ represents that correlation between ϵ_i and η_i described above,[28] and $\rho \neq 0$ represents lack of exogeneity. The key result that I derive below is that under this assumption, and letting Φ and ϕ be the cdf and pdf of the standard normal:

$$\mathbb{E}(\epsilon_i|X_i, T_i) = W_i\sigma\rho\frac{\phi(\lambda + \delta^T X_i)}{\Phi(W_i(\lambda + \delta^T X_i))} \qquad \text{for } W_i = 2T_i - 1 \tag{12}$$

notice this term is larger if the correlation ρ is larger and 0 if $\rho = 0$. Before we look at where this comes from, note that we use this for the following **two-step procedure**:

1. Fit a Probit model of $P(T_i = 1|X)$ to obtain estimates $\hat{\lambda}$ and $\hat{\delta}$
2. Use these to calculate

$$\hat{g}_i(X_i, T_i) = W_i\frac{\phi(\hat{\lambda} + \hat{\delta}^T X_i)}{\Phi(W_i(\hat{\lambda} + \hat{\delta}^T X_i))} \text{ for } W_i = 2T_i - 1$$

   for each i. This is Equation 12 without the ρ, σ, which we do not know but which we can view as the coefficient $\kappa = \sigma\rho$ in Equation 11
3. Regress Y_i on T_i, X_i and $\hat{g}_i(X_i, T_i)$ and plug into the regression or imputation estimators of section 7.2 as usual to estimate an ATE, ATC, or ATT (add a treatment-covariate interaction $\Delta^T X_i T_i$ for the imputation version).

   Tip: keep in mind that if you want to predict $\hat{Y}_i(1)$ by plugging in $T_i = 1$ into this regression, you need to plug in $T_i = 1$ both in the βT_i part and in the $\hat{g}$ expression (switching the W_i).
   Note also that we can estimate σ^2 by regressing Y_i on just X_i, T_i and use this to back out an estimate of the ρ term.

7.4.2 Limitations and Bigger Picture

The Heckman approach above depends on the bivariate normal and Probit model assumptions. If those assumptions hold, we have kind of 'magical' ability to bring in a covariate we did not actually observe, but the models may well not hold and then the above is not valid. Overall, this is a **identification strategy** (for causal effects) that relies on *parametric assumptions*, not just experimental design.

In fact, the Heckman approach, though popular in the 1980s and 1990s, is no longer used much because researchers have concerns about its validity. We introduce it above **not** as a recommendation to use it but to showcase the general idea of specifying *some* selection model and using it to adjust for selection bias. The general idea of finding a variable which, when adjusted for, makes treatment exogeneous, is present in more modern methods such as the **control function method**, which relies on ideas related to **instrumental variables** to get non-parametric identification.[29]

7.4.3 Proof of Equation 12 for $T_i = 1$

For $T_i = 1$, the goal is to show that under the model above, $\mathbb{E}(\epsilon_i|X_i, T_i = 1) = \sigma\rho\frac{\phi(\delta^T X_i)}{\Phi(\delta^t X_i)}$ wherefor simplicity of notation, I let $\lambda = 0$ (it can be added in throughout).

1. By standard results for conditional distributions of multivariate normals $\epsilon_i|\eta_i, X_i \sim N(\sigma\rho\eta_i, \sigma^2(1-\rho^2))$
2. Using the Law of Total Expectation, and letting $W_i = I(\eta_i > -\delta^t X_i) = I(T_i = 1)$

$$\begin{aligned}
\mathbb{E}(\epsilon_i|X_i, W_i = 1) &= \mathbb{E}(\mathbb{E}(\epsilon_i|\eta_i, X_i, W_i = 1)|X_i, W_i = 1) \\
&= \mathbb{E}(\mathbb{E}(\epsilon_i|\eta_i, X_i)|X_i, W_i = 1) \qquad (W_i \text{ is a function of } \eta_i, X_i) \\
&= \mathbb{E}(\sigma\rho\eta_i|X_i, W_i = 1) \\
&= \sigma\rho\int_{-X_i^T\delta}^{\infty} \eta_i\frac{\phi(\eta_i)}{\Phi(\delta^t X_i)}d\eta_i \qquad (\text{applying } W_i = 1, \text{ implies truncated normal pdf}) \\
&= \frac{\sigma\rho}{\Phi(\delta^t X_i)}\int_{-X_i^T\delta}^{\infty} \eta_i\phi(\eta_i)\eta_i \\
&= \sigma\rho\frac{\phi(-X_i^T\delta)}{\Phi(-X_i^T\delta)}
\end{aligned}$$

Where the last line follows from the fact that for fixed c,

$$\int_c^{\infty} z\phi(z)dz = \int_c^{\infty}\frac{z}{\sqrt{2\pi}}e^{-z^2/2}dz = \frac{1}{\sqrt{2\pi}}\int_{c^2/2}^{\infty} e^{-u}du = \frac{1}{\sqrt{2\pi}}(e^{-c^2/2}) = \phi(c)$$

A similar sequence of steps with some sign changes would work for $T_i = 0$ case. See also Module 7 stat 286 Review Questions, Question 3.

[28] Recall correlation is covariance divided by square root of variances so here that is $\frac{\rho\sigma}{\sigma * 1} = \rho$.
[29] See an overview in the Lecture Video

7.5 Bounds Analysis: Partial Identification

The previous sections were still framed as much of traditional causal inference and statistics is: in terms of **point identification** of some quantity (such as the ATE) for which we then obtain an estimate and ideally, a confidence interval. The catch is that obtaining that point estimate always requires some concoction of assumptions – things like randomization, unconfoundedness, normality, independence... The more assumptions we make, the more brittle our inferences may become.

Partial identification takes a step back and asks, "Do we really need an exact estimate, or would a range be enough?" and "If we made no or only minimal assumptions, could we at least rule out certain values of the quantity of interest so that we can obtain some interval or bound on it"? It is based on the idea of the **Law of Decreasing Credibility** where the more assumptions we make, the less credible our inferences become and the more people will tend to contest the plausibility of assumptions. If instead, we can restrict to a minimal set of very plausible assumptions, perhaps we can make more credible (though less specific) claims.

For example, in the tutoring experiment, you might not really care too much whether tutoring raises math scores by an average of 1, 2, 3.14159, or 4 points – it would be nice to know that, but the key question might be: does it raise scores at all? If we could obtain a **bound** on the average treatment effect that was all positive, this might be enough to decide to continue the tutoring program (or not if the bounds indicate all negative values). Another way of thinking about this is in terms of **level of resolution** – we do not always need an exact value to get useful information.

Partial identification tends to start with the 'no assumptions' scenario and work by building them up and examining how the bounds change. The kinds of assumptions it deals with tend to be different from typical statistical ones (parametric distributions, independence etc.). They tend to be substantive ones such as "what if we assume treatment never causes a decrease in outcome?"

Partial identification is closely related to sensitivity analysis. In sensitivity analysis, we might consider a "plausible range" for some sensitivity parameter (such as the unobservable partial R^2 in the previous section). One way to think about partial identification is that we instead take the full range of possible values of the sensitivity parameter so that we are no longer assuming anything about it and then examine the full range of possible values of our estimand. In partial identification, we only rule out values that are logically impossible, as will become clearer in the examples below.

7.5.1 Example 1: ATE for binary outcome and treatment under no (well, few) assumptions

If we observe some data on a binary treatment and outcome, what is the largest that the average treatment effect can be? What is the smallest? Suppose we make no only the assumptions of consistency $Y_i = Y_i(T_i)$, i.i.d. data without measurement error, and of course that the variables involved are actually binary. Then by the law of total probability,

$$\mathbb{E}(Y_i(1)) = \mathbb{E}(\mathbb{E}(Y_i(1)|T_i)) = \Pr(Y_i(1) = 1|T_i = 1)P(T_i = 1) + \Pr(Y_i(1) = 1|T_i = 0)P(T_i = 0)$$
$$\mathbb{E}(Y_i(0)) = \mathbb{E}(\mathbb{E}(Y_i(0)|T_i)) = \Pr(Y_i(0) = 1|T_i = 1)P(T_i = 1) + \Pr(Y_i(0) = 1|T_i = 0)P(T_i = 0)$$

The marginal probabilities $P(T_i = t)$ are identifiable from observation. If we assume consistency, then we can identify $\Pr(Y_i(1) = 1|T_i = 1) = \Pr(Y_i = 1|T_i = 1)$ and $\Pr(Y_i(0) = 1|T_i = 0) = \Pr(Y_i = 1|T_i = 0)$ from our data (again: not assuming any randomization here!) but the other terms in orange above are not identifiable. So what can we do? We know that probabilities are between 0 and 1 so $\mathbb{E}(Y_i(1))$ will be maximized if $\Pr(Y_i(1) = 1|T_i = 0) = 1$ and minimized if it is 0. The same applies for $\Pr(Y_i(0) = 1|T_i = 1)$ in $\mathbb{E}(Y_i(0))$. Now consider the average treatment effect $\mathbb{E}(Y_i(1)) - \mathbb{E}(Y_i(0))$. If we make the first term as large as it can be and the second as small as it can be, we have the maximum average treatment effect and if we flip this, we have the minimum average treatment effect.

$$\tau_{max} = \Pr(Y_i(1) = 1|T_i = 1)P(T_i = 1) + \Pr(T_i = 0) - (\ \Pr(Y_i(0) = 1|T_i = 0)P(T_i = 0)\ )$$
$$\tau_{min} = \Pr(Y_i(1) = 1|T_i = 1)P(T_i = 1) - \ \Pr(Y_i(0) = 1|T_i = 0)P(T_i = 0) - \Pr(T_i = 1)$$

If we take the difference, we get

$$\tau_{max} - \tau_{min} = \Pr(T_i = 0) + \Pr(T_i = 1) = 1$$

What does this tell us? It tells us that although a priori, τ can take any value in $[-1, 1]$, once we learn the observable (or at least estimable – ignore estimation error for now) quantities, $P(T_i = 1), P(T_i = 0), P(Y_i = 1|T_i = 1), P(Y_i = 1|T_i = 0)$, these impose some constraints in the sense that they isolate the possible average treatment effect to a range of length 1. Exactly what that range will be depends on our data. For example, suppose we have a balanced set-up with $P(T_i = 0) = P(T_i = 1) = \frac{1}{2}$ and suppose we know $\Pr(Y_i = 1|T_i = 1) = .7$ and $\Pr(Y_i = 1|T_i = 0) = .3$. Then the above tells us that

$$\tau_{max} = (.7)(.5) + .5 - (.3)(.5) = .7$$
$$\tau_{min} = (.7)(.5) - (.3)(.5) - .5 = -.3$$

So given this information, we know the average treatment effect is in $[-.3, .7]$. That might not seem helpful, but we know this having made no assumptions about randomness or unconfoundedness!

7.5.2 Example 2: binary treatment, continuous outcome, with self-selection

Bounds analyses can also be used to explore what would happen in scenarios where assumptions like unconfoundedness are violated. For example, suppose we have a binary treatment and continuous outcome and suppose we have a violation of randomized treatment in the form of self-selection into better outcomes. That is, supposing that higher outcomes are better, assume that if unit i has $Y_i(1) > Y_i(0)$ then unit i gets $T_i = 1$ and vice versa. This implies that:

$$\mathbb{E}(Y_i(0)|T_i = 1) \leq \mathbb{E}(Y_i(1)|T_i = 1) = \mathbb{E}(Y_i|T_i = 1) \tag{13}$$

$$\mathbb{E}(Y_i(1)|T_i = 0) \leq \mathbb{E}(Y_i(0)|T_i = 0) = \mathbb{E}(Y_i|T_i = 0) \tag{14}$$

Instead of a unit-level dynamic, we could also start by assuming the inequalities above as an on-average dynamic. Suppose also that outcome Y_i has a lower bound l.

If we again use the Law of Total Probability, we get non-identifiable quantities[30] $\mathbb{E}(Y_i(1)|T_i = 0)$ and $\mathbb{E}(Y_i(0)|T_i = 1)$. We can obtain maximum and minimum values of $\mathbb{E}(Y_i(1)), \mathbb{E}(Y_i(0))$ by setting these to their observational upper bounds from Equations 13 and 14 or to the lower bound l.

$$\begin{aligned}
\mathbb{E}(Y_i(1)) &= \mathbb{E}(Y_i(1)|T_i = 1)P(T_i = 1) + \mathbb{E}(Y_i(1)|T_i = 0)P(T_i = 0) \\
&\leq \mathbb{E}(Y_i(1)|T_i = 1)P(T_i = 1) + \mathbb{E}(Y_i(0)|T_i = 0)P(T_i = 0) \\
&= \mathbb{E}(Y_i|T_i = 1)P(T_i = 1) + \mathbb{E}(Y_i|T_i = 0)P(T_i = 0) \\
&= \mathbb{E}(Y_i)
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}(Y_i(1)) &\geq \mathbb{E}(Y_i(1)|T_i = 1)P(T_i = 1) + lP(T_i = 0) \\
&= \mathbb{E}(Y_i|T_i = 1)P(T_i = 1) + lP(T_i = 0)
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}(Y_i(0)) &= \mathbb{E}(Y_i(0)|T_i = 1)P(T_i = 1) + \mathbb{E}(Y_i(0)|T_i = 0)P(T_i = 0) \\
&\geq lP(T_i = 1) + \mathbb{E}(Y_i|T_i = 0)P(T_i = 0)
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}(Y_i(0)) &\leq \mathbb{E}(Y_i(1)|T_i = 1)P(T_i = 1) + \mathbb{E}(Y_i(0)|T_i = 0)P(T_i = 0) \\
&= \mathbb{E}(Y_i|T_i = 1)P(T_i = 1) + \mathbb{E}(Y_i|T_i = 0)P(T_i = 0) \\
&= \mathbb{E}(Y_i)
\end{aligned}$$

Using these we get $\tau \in [\tau_{max}, \tau_{min}]$

$$\tau_{max} = \mathbb{E}(Y_i) - lP(T_i = 1) - \mathbb{E}(Y_i|T_i = 0)P(T_i = 0)$$

$$\tau_{min} = \mathbb{E}(Y_i|T_i = 1)P(T_i = 1) + lP(T_i = 0) - \mathbb{E}(Y_i)$$

Again, ignoring estimation error, we could calculate these bounds because they are based on observational quantities. For example, if outcomes are math scores, we have lower bound 0. Suppose that observationally, $\mathbb{E}(Y_i|T_i = 1) = .80$ and $\mathbb{E}(Y_i|T_i = 0) = .60$ and $\mathbb{E}(Y_i) = .75$ with $P(T_i = 1) = .40$. Then we know that in the situation where we assume units self-select to better treatment, the true treatment effect must be within

$$\tau \in [.80(.40) - .75, .75 - .60(.60)] = [-.40, .39]$$

It is interesting to note that despite the treatment group mean being clearly larger than the control group mean, under self-selection, the data are consistent with a fairly large average negative effect $\mathbb{E}(Y_i(1) - Y_i(0))$. Note that there is nothing here that makes any value in the interval above more likely than the other. We would not, for example, take the middle of the interval as a point estimate. Given the assumptions we made, **we only say that all values in the interval are compatible with the observations.**

7.5.3 Example 3: Instrumental Variables with Binary Instrument and Outcome Manski (1990)

In the standard IV set-up, we can only point identify the complier ATE. Can we say anything about the population ATE? Let's consider the standard IV set-up, assuming **randomization** of instrument Z_i, the **exclusion restriction**, and **monotonicity** as usual and see what we can see about the population ATE. To simplify notation, let $\mu_{tz} = \Pr(Y_i = 1|T_i = t, Z_i = z)$ and $\pi_z = \Pr(T_i = 1|Z_i = z)$ (observational quantities). By **randomization** of the instrument,

[30] Unidentifiable without further assumptions like randomization

$$\mathbb{E}(Y_i(t)) = Pr(Y_i(t) = 1) = Pr(Y_i(t) = 1|Z_i = t) \qquad \text{for } t = 0, 1 \tag{15}$$

Implicitly here we are also using the exclusion restriction so that $Y_i(t) = Y_i(t, z)$ for $z = 0, 1$ and hence it makes sense to speak of simply $\mathbb{E}(Y_i(t))$ and not separate $\mathbb{E}(Y_i(t, z))$ quantities. Using the law of total probability and again highlighting terms that emerge which are non-identifiable, we then have:

$$\begin{aligned}\mathbb{E}(Y_i(1)) = Pr(Y_i(1) = 1|Z_i = 1) &= Pr(Y_i(1) = 1|T_i = 1, Z_i = 1)\pi_1 + Pr(Y_i(1) = 1|T_i = 0, Z_i = 1)(1 - \pi_1) \\ &= \mu_{11}\pi_1 + Pr(Y_i(1) = 1|T_i = 0, Z_i = 1)(1 - \pi_1)\end{aligned}$$

$$\begin{aligned}\mathbb{E}(Y_i(0)) = Pr(Y_i(0) = 1|Z_i = 0) &= Pr(Y_i(0) = 1|T_i = 1, Z_i = 0)\pi_0 + Pr(Y_i(0) = 1|T_i = 0, Z_i = 0)(1 - \pi_0) \\ &= Pr(Y_i(0) = 1|T_i = 1, Z_i = 0)\pi_0 + \mu_{00}(1 - \pi_0)\end{aligned}$$

We can again obtain an upper bound and lower bound by setting the non-identified quantities to (1,0) and (0,1).

$$\tau_{min} = \mu_{11}\pi_1 - 1(1 - \pi_1) - \mu_{00}(1 - \pi_0) \qquad \tau_{max} = \mu_{11}\pi_1 + 1\pi_0 - \mu_{00}(1 - \pi_0)$$

The width of this interval is $\tau_{max} - \tau_{min} = 1 - (\pi_1 + \pi_0)$. Notice that $\pi_1 + \pi_0 = \Pr(T_i = 1|Z_i = 1) - P(T_i = 1|Z_i = 0) = \Pr(T_i(1) = 1) - \Pr(T_i(0) = 1) = \mathbb{E}(T_i(1) - T_i(0))$ is exactly the proportion of compliers! (see Module 5 for more) Intuitively, then, if everyone is a complier, we $\tau_{max} - \tau_{min} = 0$ and have point identification of the ATE. If many are compliers, then the observed information narrows the bounds a lot, and if few are compliers, the observed information yields broad bounds on the ATE. Notice that this argument did not incorporate monotonicity, so it actually applies even if we do not assume monotonicity.

In fact, the bounds above are not **sharp**. Bounds are sharp if they are the tightest (shortest interval) possible given the information available. The approach above does not fully use the information available. In Equation 15, we condition on $Z_i = 1$ or $Z_i = 0$ immediately and end up considering four types of units in terms of their observed $(T_i = t, Z_i = z)$ combinations. In fact, for a binary outcome, under the exclusion restriction, there are $2^4 = 16$ types of people, defined by $U_i := (T_i(0), T_i(1), Y_i(0), Y_i(1))$ values.[31] For example, a $(0, 0, 1, 1)$ is a never-taker who always has outcome 1 (e.g., never gets tutoring, always passes the math test) and $(0, 0, 0, 1)$ is a never-taker who passes only if treated (though we never observe the $Y_i(1) = 1$). If we also add on the monotonicity assumption, we remove defiers and end up with 12 types. In either case, suppose we have defined a set $\mathcal{U}$ of possible types. The causal quantity can be broken down as:

$$\Pr(Y_i(1) = 1) - \Pr(Y_i(0) = 1) = \sum_{u \in \mathcal{U}} (\Pr(Y_i(1) = 1|U_i = u) - \Pr(Y_i(0) = 1|U_i = u)) \Pr(U_i = u) \tag{16}$$

$\Pr(Y_i(t) = 1|U_i = u)$ are just 0 or 1, but $\Pr(U_i = u)$ is unknown. However, we can formulate this as a **linear programming** problem where we search over values of $\Pr(U_i = u)$ to minimize and maximize 16 subject to the constraints:

1. $\Pr(U_i = u) \in [0, 1]$ for all $u \in \mathcal{U}$
2. $\sum_{u \in \mathcal{U}} \Pr(U_i = u) = 1$
3. Consistent with our observed data,

$$\begin{aligned}\Pr(Y_i = y, T_i = t, Z_i = z) &= \sum_{u \in \mathcal{U}} \Pr(Y_i = y, T_i = t, Z_i = z, U_i = u) \\ &= \sum_{u \in \mathcal{U}} \Pr(Y_i(t) = y, T_i(z) = t|Z_i = z, U_i = u)\Pr(U_i = u|Z_i = z) \Pr(Z_i = z) \\ &= \sum_{u \in \mathcal{U}} \Pr(Y_i(t) = y, T_i(z) = t|U_i = u)\Pr(U_i = u) \Pr(Z_i = z) \qquad \text{(can drop } Z_i = z \text{ because randomization)}\end{aligned}$$

where above, we are using the assumption that Z_i is randomized so that $\Pr(U_i = u|Z_i = z) = \Pr(U_i = u)$ and the fact that once we condition on $U_i = u$ and $Z_i = z$, we know $T_i(z)$ and $Y_i(t)$. The leftmost probability within the sum is essentially just an indicator for whether the specified y, t, z configuration aligns with the type $U = u$. For example, for the $u = (0, 1, 0, 1)$ type (a complier whose outcome is 1 when treated and 0 when not treated), we know that $\Pr(Y_i(1) = 1, T_i(0) = 1|U_i = u) = 1$ and $\Pr(Y_i(1) = 0, T_i(0) = 1|U_i = u) = 0$.

These results are due to Balke and Pearl (1997).

[31] Note that this is $T_i(z)$ for $z = 0, 1$ and then $Y_i(t)$ for $t = 0, 1$. It is $Y_i(t)$ and not $Y_i(t, z)$ because of the exclusion restriction. See Section 5.2. In fact, if we were to *not* assume the exclusion restriction, we would have 2^6 hypothetical types corresponding to $T_i(z)$ for $z = 0, 1$ and $Y_i(t, z)$ for all combinations. We could do a bounds analysis in the no-exclusion case by optimizing over this larger set $\mathcal{U}$.

7.5.4 Estimating Bounds

In the discussion above, the "observational" quantities were only observational in an identification sense. We have to estimate them in practice. That is, the *bounds* themselves are now our estimands and we should think about how well we can estimate them and how to quantify their uncertainty. There are at least two possible approaches to creating confidence intervals here. The first deals only with τ_{min}, τ_{max} as estimands. The second recognizes that our true estimand is still τ (we just might not have enough assumptions to point identify it) and focuses on obtaining an interval which, at least asymptotically, contains τ at least $(1-\alpha)$ of the time. In both cases, we **assume** it is possible to argue that $\hat{\tau}_{min}, \hat{\tau}_{max}$ are asymptotically normal so that standard normal approximation confidence intervals apply.

1. **Confidence intervals for true bounds τ_{min}, τ_{max} (conservative)**
   Quantities like $P(T_i = t)$, $P(Y_i = y|T_i = t)$ or $\mathbb{E}(Y_i|T_i = 1)$ have unbiased estimators that take the form of sample means, so we can estimate their variances and apply the CLT to obtain standard normal approximation based asymptotic confidence intervals. If needed, we can also apply the multivariate CLT to argue they are jointly normal and the Delta Method to calculate the asymptotic variance of $\hat{\tau}_{min}$ or $\hat{\tau}_{max}$, which are usually some product or sum of these quantities. Given $\hat{\sigma}_{min}$, an estimate of the standard error of the lower bound estimator $\hat{\tau}_{min}$ and $\hat{\sigma}_{max}$, an estimate of the standard error of the upper bound estimator, we can obtain a two-sided $(1-\alpha)\%$ confidence interval and combine them to extend our bound as follows:

$$C_1 = \left[\hat{\tau}_{min} - z_{1-\alpha/2}\hat{\sigma}_{min} \ , \ \hat{\tau}_{max} + z_{1-\alpha/2}\hat{\sigma}_{max}\right]$$

   Note: we should always still algebraically get $\hat{\tau}_{min} \leq \hat{\tau}_{max}$ since we estimate the bounds using empirical probabilities from the same data.
   For this interval, proper coverage would mean that C_1 contains the true *bounds* $(1-\alpha)$ of the time. The interval above achieves at least this coverage because, letting U and L represent upper and lower bounds of the confidence interval,[32]

$$\begin{aligned}
\Pr(L \leq \tau_{min}, \tau_{max} \leq U) &\geq \Pr(\tau_{min} \geq L) + \Pr(\tau_{max} \leq U) - 1 \\
&= (1 - \frac{\alpha}{2}) + (1 - \frac{\alpha}{2}) - 1 \\
&= 2 - \alpha - 1 \\
&= 1 - \alpha
\end{aligned}$$

   where here we use $P(A \cup B) = P(A) + P(B) - P(A \cap B) \geq P(A) + P(B) - 1$ and we use the fact that for a two-sided $1-\alpha$ confidence interval, the probability of being greater than just the lower bound is $1 - \alpha/2$ and similarly for being lower than just the upper bound.

   Example: σ_{min} variance calculation assuming i.i.d. sample
   Consider $\tau_{min} = \mathbb{E}(Y_i|T_i = 1)P(T_i = 1) + lP(T_i = 0) - \mathbb{E}(Y_i)$ from Example 2. A consistent estimator of this is, using Law of Large Numbers and Slutsky,

$$\begin{aligned}
\hat{\tau}_{min} &= \left(\frac{1}{\sum_{i=1}^n T_i} \sum_{i=1}^n T_i Y_i\right) \bar{T} + l(1 - \bar{T}) - \bar{Y} \\
&= \frac{1}{n} \sum_{i=1}^n T_i Y_i + l(1 - \bar{T}) - \bar{Y} \\
&= \frac{1}{n} \sum_{i=1}^n (T_i Y_i + l(1 - T_i) - Y_i)
\end{aligned}$$

   By the CLT, this is asymptotically normal with asymptotic variance

$$\mathbb{V}(T_i Y_i + l(1 - T_i) - Y_i) = \mathbb{V}(T_i Y_i) + l^2 \mathbb{V}(1 - T_i) + \mathbb{V}(Y_i) - 2\text{Cov}(T_i Y_i, Y_i) + 2l\text{Cov}(T_i Y_i, 1 - T_i) - 2l\text{Cov}(1 - T_i, Y_i)$$

   Although this is an ugly expression, each quantity here is estimable via finite-sample variances and covariances. We will ultimately then have

$$\hat{\sigma}_{min} = \frac{\sqrt{\hat{\mathbb{V}}(T_i Y_i + l(1 - T_i) - Y_i)}}{\sqrt{n}}$$

   (by the usual logic that for i.i.d. random variables Y_i, the variance of a sample mean is of form σ^2/n with $\sigma^2 = \mathbb{V}(Y_i)$)

[32] These are random variables, while τ_{min}, τ_{max} are fixed true values.

2. **Confidence intervals for true value of estimand τ (preferred method)**

The previous approach has a drawback. Yes, we have that it will contain the *true bounds* τ_{min}, τ_{max} with the right probability, but really, we are interested in whether it will contain the true τ that proportion of the time. It turns out that if we consider this question directly, we can get a narrower confidence interval that still asymptotically has the right coverage for the **true** τ.

Consider two one-sided $1-\alpha$ confidence intervals, one for $\hat{\tau}_{min}$ and one for $\hat{\tau}_{max}$:

$$\begin{aligned} \text{For } \hat{\tau}_{max}: &\quad (-\infty, \hat{\tau}_{max} + z_{1-\alpha}\hat{\sigma}_{max}] \\ \text{For } \hat{\tau}_{min}: &\quad [\hat{\tau}_{min} - z_{1-\alpha}\hat{\sigma}_{min}, \infty) \end{aligned}$$

If we then take the intersection, we get

$$C_2 = [\hat{\tau}_{min} - z_{1-\alpha}\hat{\sigma}_{min} \ , \ \hat{\tau}_{max} + z_{1-\alpha}\hat{\sigma}_{max}]$$

which is narrower than C_1 above since for $\alpha < .5$, $z_{1-\alpha} \leq z_{1-\alpha/2}$. The intuitive justification for using these one-sided intervals is as follows: If the true τ is less than the true τ_{max}, then if $\hat{\tau}_{max}$ is consistent for τ_{max}, it will always eventually be greater than τ so that the estimated interval contains τ from above with probability approaching 1. However, if it happens that $\tau = \tau_{max}$, then we have to be careful because if, for example, $\hat{\tau}_{max}$ converges towards τ_{max} from below, then our estimated interval $[\hat{\tau}_{min}, \hat{\tau}_{max}]$ never contains the true value! Hence we widen our interval by $z_{1-\alpha}\hat{\sigma}_{max}$ above (a quantity which will converge towards 0 as $n \to \infty$ but always adds a little bit extra to the width of our estimated interval). The same logic applies to the lower bound, with $\tau_{min} = \tau$ as a special case.

Formally, we have:

$$\Pr(\tau \geq L, \tau \leq U) = \Pr(\hat{\tau}_{min} - z_{1-\alpha}\hat{\sigma}_{min} \leq \tau \leq \hat{\tau}_{max} + z_{1-\alpha}\hat{\sigma}_{max}) \to \begin{cases} 1 & \text{if } \tau \in (\tau_{min}, \tau_{max}) \\ 1-\alpha & \text{if } \tau = \tau_{max} \text{ or } \tau = \tau_{min} \end{cases}$$

This results from the following logic:

- Suppose $\tau_{min} < \tau < \tau_{max}$. Then if $\hat{\tau}_{min}$ and $\hat{\tau}_{max}$ are consistent (converge in probability to true values), the probability that the interval they define contains τ converges to 1
- Suppose $\tau = \tau_{max}$. Then as long as $\hat{\tau}_{min}$ is consistent, it will eventually be less than τ with probability 1 so with probability $\to 1$, the interval contains the true τ from below. However, from above, by construction of the one-sided confidence interval for τ_{max}, the coverage of $\tau = \tau_{max}$ is $1-\alpha$.
- The same as the previous happens if $\tau = \tau_{min}$.

More formally, suppose for simplicity, we know σ_{min} and σ_{max} and suppose $\tau = \tau_{max}$. Then, again using $\Pr(A \cap B) \geq P(A) + P(B) - 1$,

$$\begin{aligned} \Pr(\tau \in C_2) &= \Pr(\hat{\tau}_{min} - z_{1-\alpha}\sigma_{min} \leq \tau_{max} \leq \hat{\tau}_{max} + z_{1-\alpha}\sigma_{max}) \\ &\geq \Pr(\hat{\tau}_{min} - z_{1-\alpha}\sigma_{min} \leq \tau_{max}) + \Pr(\tau_{max} \leq \hat{\tau}_{max} + z_{1-\alpha}\sigma_{max}) - 1 \\ &= 1 + o(1) + \Pr(\tau_{max} \leq \hat{\tau}_{max} + z_{1-\alpha}\sigma_{max}) - 1 \\ &= \Pr\left(-z_{1-\alpha} \leq \frac{\hat{\tau}_{max} - \tau_{max}}{\sigma_{max}}\right) + o(1) \\ &\to \Phi(z_{1-\alpha}) = 1-\alpha \end{aligned}$$

Where the last line follows assuming we can show asymptotic normality and the $1 + o(1)$ follows from the following argument, which holds as long as $\tau_{max} > \tau_{min}$ (else we have point identification and can make usual confidence interval arguments),

$$\Pr(\hat{\tau}_{min} \leq \tau_{max} + z_{1-\alpha}\sigma_{min}) \geq \Pr(\hat{\tau}_{min} \leq \tau_{max}) = \Pr(\hat{\tau}_{min} - \tau_{min} \leq \tau_{max} - \tau_{min}) \to 1$$

because by convergence in probability of $\hat{\tau}_{min}$, $\Pr(|\hat{\tau}_{min} - \tau_{min}| \leq \epsilon) \to 1$ for any $\epsilon > 0$ and hence $\Pr(\hat{\tau}_{min} - \tau_{min} \leq \epsilon) \to 1$.

General note: Usually, we expect a confidence interval to shrink in width to 0 as $n \to \infty$. Here, that does not happen. The confidence interval should converge to the true interval $[\tau_{min}, \tau_{max}]$

7.6 Module 7 - Part II: DAGs (a brief introduction)

The Directed Acyclic Graph (DAG) and graphs in general are widely used tools in statistics for representing complex dependence structures among different random variables. Given a graph, there are formal rules for reading off the conditional independences it implies and writing down a joint distribution. **Causal** DAGs refer to the case within the broader subject of 'graphical models' when the relationships implied by a graph are **assumed** to be causal. These DAGs come with some special rules and operations for working with them, including the **do calculus**. DAGs can be extremely useful for visualizing and thinking through causal assumptions and for figuring out, given a complex causal structure, which quantities can be identified and via what conditioning. Note, however, that in general, the causal structure in a causal DAG is *assumed*, not discovered from the data, and hence the resulting identification results are only valid given the structure.[33] The DAG flavor of causal inference literature is sometimes criticized for focusing on this identification perspective without enough attention to the difficulty of specifying a causal structure in the first place and the issues of estimation and inference the follow identification (Imbens, 2020, p. 5-7).

How do DAGs relate to potential outcomes? Potential outcomes and DAGs represent two different traditions that do not necessarily conflict but sometimes have trouble communicating. Unfortunately, there is no easy link between the two perspectives [34] and some kinds of assumptions can be harder to represent (e.g., harder to talk about principal strata and their relevance for identification as in module 5, harder to represent treatment effect heterogeneity). This course focuses on the potential outcomes perspective, but this section of the module is meant to give you a brief taste of the other main flavor of causal thinking. Where possible, we note connections to potential outcomes.

7.6.1 DAG Basics

In general, a **graph** G is defined by a set of **vertices** V and a set of **edges** E connecting them. For a Directed Acyclic Graph (DAG) these edges are **directed**, meaning the edges are represented as arrows, and we assume the graph is **acyclic**, meaning there are no cyclical structures of the form $A \to B \to C \to A$ that would indicate simultaneity (A causing B but B also causing A). In the causal DAG context, the nodes represent variables, with solid lines generally representing variables that are observed and dashed lines representing ones that are unobserved (U in Figure 16). The arrows represent causal relationships (more below) and importantly, the lack of an arrow represents a lack of direct causal relationship. For example, in Figure 16, T has a direct causal effect on Y but X does not because there is no $X \to Y$ arrow. But does it follow that if we **intervene** to set the value of X (e.g., in the context of an experiment), there will be no effect on Y? No! If we intervene on X, the DAG tells us that this would affect T, which would affect Y. Moreover, intervening on X would also affect Z, which affects T, which affects Y! In this way, DAGs can represent complicated causal relationships, but that complexity also means we need some formal ways of reading off causal relationships from DAGs – more on that below.

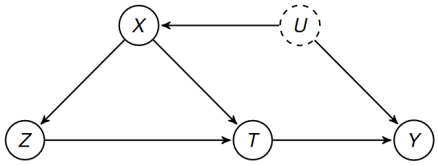


Figure 16: Example DAG taken from Slide 2.

First, some further graph terminology that will be useful:

- The **parents** of a node j are all vertices i with an arrow $i \to j$ with no intermediaries and similarly, if $i \to j$, then j is a **child** of i. Two nodes with an arrow between them have a **direct** causal relationship.[35] Above, T is a parent of Y. The set of parents of a node i is often denoted $pa(i)$.
- The **ancestors** of node j are all the nodes with a **directed path** $i \to ... \to j$ that leads to i. We also say i is a **descendant** of j. Above, Z, X, U are all ancestors of T. If the arrow $X \leftarrow U$ instead pointed $X \to U$, then U would not be an ancestor of T.
- A **path** is any sequence of nodes with some link between them, regardless of which ways the arrows point. A **causal path** or **chain** is a path where the arrows do all point in the same direction. Above, $X \to T \to Y$ is a causal path while $X \leftarrow U \to Y$ is a path but not a causal one.

Notation note: Usually, V is defined as a indices $\{1, ..., n\}$ where index i represents a random variable X_i and X_V represents the vector $(X_1, ..., X_n)$. Below, in particular examples, I refer to nodes directly by their random variables (T, U, X etc.) but for more general statements, I use the X_i notation. For example, $X_{pa(i)}$ and $pa(X_i)$ mean the same thing.

[33] There is line of work on *causal discovery* that we do not cover in this course

[34] Some have proposed a methods of linking the two called Single World Intervention Graphs (SWIG) (see here and here), but these have not caught on widely. See also Imbens (2020) for an essay comparing and contrasting the two approaches.

[35] Hidden in any arrow are mechanisms (e.g., the biological reasons that ibuprofen might lessen pain) but the single arrow between X and T in Figure 16 indicates that we assume none of these mechanisms directly affect or are directly affected by the other variables.

7.6.2 Reading DAGs

Factorization of Log Likelihood

A primary purpose of a DAG is to encode conditional independences. A key type of conditional independence it specifies is the following: given vertex set $V = \{1, ..., J\}$ indexing random variables $X_1, ..., X_J$ and given and $i \in V$ and letting $pa(i)$ be the parent nodes of i and $V \backslash pa(i)$ represent all that are not in $pa(i)$

$$X_i \perp\!\!\!\perp X_{V \backslash pa(i)} | X_{pa(i)} \tag{17}$$

That is, variable X_i is independent of all non-parent variables conditional on its parents. This independence holds for any set of random variables with a joint distribution consistent with the DAG. It implies that the overall joint distribution of $(X_1, ..., X_J)$ factorizes as follows:

$$P(X_1, ..., X_J) = \prod_{j=1}^{J} P(X_j | pa(X_j))$$

In Figure 16, this is

$$P(T, U, X, Y, Z) = P(Y|T, U)P(T|X, Z)P(Z|X)P(X|U)P(U)$$

DAG as non-parametric SEM

Recall that a **Structural Equation Model** (Section 4.1) is one that poses causal relationships instead of only statistical (association-based) ones. We previously discussed the linear structural equation model, where those causal relationships are linear, and the idea that these models *assume* causality rather than *discover it*. One way to think about a causal DAG is that it is a **non-parametric** structural equation model. It poses that there are certain causal relationships among variables, but it does not pose the functional form of these relationships. Any more specific SEM, including linear SEMs, are special cases of this general specification. One way of reading the DAG in Figure 16 is to see it a specifying the following set of relationships which follow the factorization of the joint above:

$$\begin{aligned} Y &= f_1(T, U, \epsilon_1) \\ T &= f_2(X, Z, \epsilon_2) \\ Z &= f_3(X, \epsilon_3) \\ X &= f_4(U, \epsilon_4) \end{aligned}$$

where each ϵ_j is an error term. This is non-parametric if we do not specify the form of $f_1, ..., f_4$. That includes not specifying the form or distribution of the error term. This is one way that we connect DAGs to potential outcome models (above we could write $Y(t) = f_1(t, U, \epsilon_1)$).

Reading Conditional Independences in General: d-separation

Equation 17 is actually a special case of a more general result that relies on **d-separation**. To define this, we need to define blocking and colliders (see Figure 17 for examples)

- **Definition:** A **collider** is a vertex which has two incoming arrows along a path. In Figure 16, on the path $Z \to T \leftarrow X \to U$, T is a collider. However, on the path $Z \to T \to Y$, T is not a collider so this designation is relative to the path.
- **Definition:** a path from vertex i to vertex j is **blocked** by set of vertices C if either (1) the path includes a non-collider in C or (2) the path includes a collider that is **not** in C and no descendent of any collider is in C.
- **Definition:** Let A, B, C be disjoint sets of vertices. Then A and B are ***d*-separated** by C if every path from a node $a \in A$ to a node $b \in B$ is blocked by C.

Key result: If A and B are d-separated by C, $X_A \perp\!\!\!\perp X_B | X_C$. If A and B are not d-separated (i.e., d-connected), then there exists at least one distribution compatible with the DAG such that $X_A \not\perp\!\!\!\perp X_B | X_C$.[36]

[36] Compatibility with the DAG requires that the joint distribution at least factorize as described above, but a joint distribution with more independences is also compatible. For example, if in Figure 16, there were no arrows so that T, U, X, Y, Z were all totally independent, we would still say this is compatible with the DAG. It is only if there is dependence that the DAG does not allow that we have incompatibility.

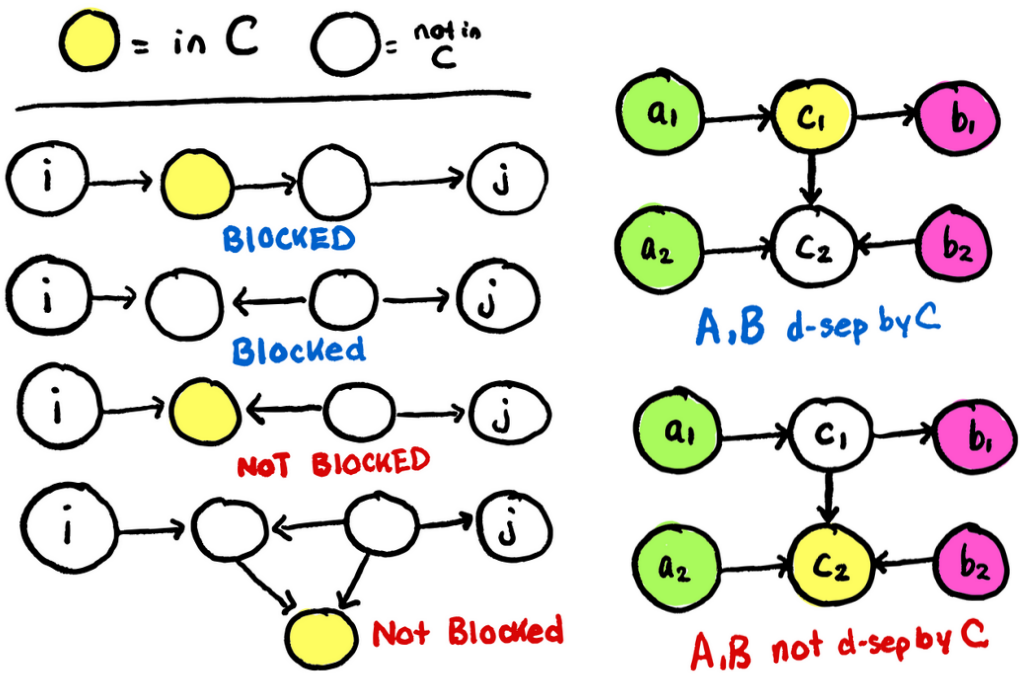


Figure 17: **Left**:Illustration of path blocking. **Right**: illustration of d-separation. In the top graph, $a_1 \to c_1 \to b_1$ is blocked because c_1 is a non-collider and is in C. All other a_i to b_j paths are blocked because they go through c_2, which is a collider on each of those paths and is not in C. In the bottom graph, not only is $a_1 \to c_1 \to b_1$ not blocked but all the other paths are not either because a collider on those paths is in C. That is, even if we added c_1 to C, the bottom graph would fail d-separation.

We can use d-separation to read off a more general set of distributional independences implied by a DAG. For example, in Figure 18, we might ask: What variables do I need to condition on for W and Y to be conditionally independent?

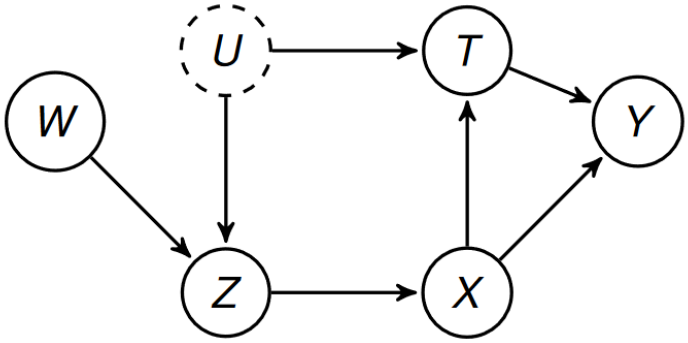


Figure 18: Example from Slide 6

First, consider the paths between them and the conditioning set $C = \emptyset$

- $W \to Z \to X \to Y$ - unblocked, only non-colliders and none conditioned on
- $W \to Z \to X \to T \to Y$ - unblocked, only non-colliders and none conditioned on
- $W \to Z \leftarrow U \to T \to Y$ - blocked, Z is a collider not conditioned on
- $W \to Z \leftarrow U \to T \leftarrow X \to Y$ - blocked, Z, T are colliders not conditioned on

From this, we see that W is not d-separated from Y by $\emptyset$ so it is possible that $W \not\perp\!\!\!\perp Y$. We also see that if we were to condition on Z, we un-block the second two paths. However, if we condition on X, this will not unblock any paths and will block the first two. We could also condition on X and T or X, T, Z but not T and Z since the fourth path would then be unblocked. That is, the following are true for any distribution compatible with this DAG:

$$W \perp Y | X \qquad\qquad W \perp Y | X, T, Z$$

7.6.3 The Backdoor Criterion and causal identification

If we assume that a DAG is causal, then directed paths in the DAG become causal and the question becomes what to condition on to achieve **unconfoundedness** (Section 7.1). The intuition for the **backdoor criterion** for causal identification of the effect of T on Y is that we find a conditioning set X which blocks all back-door (non-causal) paths between T and Y and avoids conditioning on any post-treatment variables, which we know can create bias.

Formally: a **back-door path** is any path between T and Y with an arrow incoming into T, and a **back-door adjustment set** C for T, Y is one such that

1. Every backdoor path between T and Y is blocked by X
2. X does not contain any descendants of T

Given such a set X, it is possible to show the **backdoor adjustment formula**

$$P(Y(t)) = P(Y|T = t) = \sum_{x \in \mathcal{X}} P(Y|T = t, X = x)P(X = x)$$

(we could also replace this with $\mathbb{E}(Y(t))$ and $\mathbb{E}(Y|T = t, X = x)$). This is an identification equality. The right side is observational, but the left side is causal. You might notice the above is written in potential outcomes notation. The more standard DAG literature way of writing this would be

$$P(Y|do(T)) = \sum_{x} P(Y|T = t, X = x)P(X = x)$$

You may be surprised it is not $P(X = x|T = t)$ as per the usual law of total probability. This has to do with with the fact that we are *intervening* on T. I explain both these ideas in the next section. But first, consider Figure 18. There, $T \leftarrow X \rightarrow Y$ is a back-door path for T and Y, as is $T \leftarrow U \rightarrow Z \rightarrow X \rightarrow Y$. These are both blocked by conditioning on X, which is in both cases a non-collider and not a post-treatment variable. Hence the variable X is a valid back-door adjustment set. See another example on Slide 8.

7.6.4 Why no conditioning on colliders?

Get into the causal DAG world and you will likely hear people worry about"conditioning on colliders." Above, we also said a requirement for a path to be blocked by C is that it contains no colliders. What is so bad about these variables? One way to think about this is that it is closely related to the issue of **post-treatment** bias. For example, in the case of an experiment where we randomly assign kids to tutoring, we might have the following DAG. Here, tutoring affects a math test score, which affects placement in advanced math for the following school year, but it might also affect other factors (grades on other assignments, teacher recommendation) which also affect placement. If we now condition on children placed into advanced math, we might not observe much affect of tutoring on math scores because these are all children who scored high. We destroy our ability to detect the dynamic where tutoring causes some children to be placed in advanced math who otherwise would not have been. This example also appeared in Section 4.5.1.

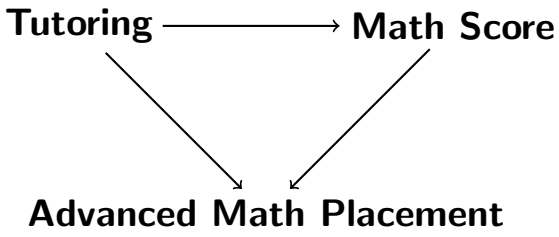


Figure 19: Example DAG for Tutoring Experiment

7.6.5 The do operation for causal DAGs

d-separation or factorization of the joint are not specific to causal DAGs. These are general principles for reading off conditional independences, causal or not. Causality in DAG world is considered via the idea of an intervention, represented by the **do operation**. $do(x)$ means " intervene on X by setting its value to x." For example, imagine we live in a world where there are genetic factors U that cause both smoking T and cancer Y but also smoking has some causal effect on cancer. This is represented by Figure 20. In this case, $do(t)$ represents us actually intervening to make people smoke ($t = 1$) or not ($t = 0$), as we would if we randomly assigned them in an experiment.

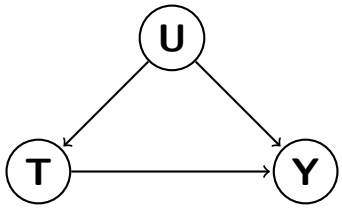


Figure 20: Example DAG where U is genetics, T is smoking, and Y is cancer.

A key property of $do(t)$ is that it breaks incoming arrows into T. That is, given $do(t)$, the DAG above becomes:

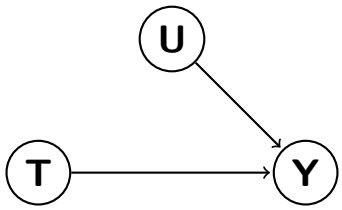


Figure 21: Example DAG after $do(t)$.

This reflects the notion that if we randomly assigned people to smoke or not, then even if there are genetic factors which would otherwise have caused them to smoke or not, now, those factors have no more effect on smoking. *We* caused them to smoke or not. The link to potential outcomes is:

$$P(Y|do(t)) = P(Y(t))$$

where $P(Y|do(1))$ is essentially "the distribution of outcomes if I forced everyone to smoke" and $P(Y|do(0))$ is "the distribution of outcomes if I forced everyone to not smoke." These are also called **counterfactual distributions**. If we have experimental data with randomized treatment, then we actually observe data from the world in the Figure 21 DAG and $P(Y|do(t)) = P(Y|T = t)$. But if we have observational data, then we actually have data from the world in the Figure 20 DAG and $P(Y|do(t)) \neq P(Y|T = t)$. The **backdoor adjustment formula** is about taking data generated by the observational world and linking it to the quantity we'd like from the intervention world. This is all essentially equivalent to the ideas we've been discussing throughout the course about linking causal quantities defined via potential outcomes to observational quantities, but now, the causal quantities are defined in terms of the do operation and the causal relationships represented by a DAG.

What does the **do(x)** operation breaking incoming arrows into X in the DAG mean in terms of writing down distributions? Suppose we have a graph with vertices indexed by $V = \{1, ..., n\}$ which, as discussed above, factorizes as

$$P(X_V) = \prod_{i=1}^{n} P(x_i|x_{pa(i)})$$

Suppose now we $do(x_j)$ for some variable x_j. The DAG representing this world drops any arrows $i \to j$, and x_j ceases to be random because we fix it (a kind of causal conditioning). The joint distribution distribution according to that updated DAG is then:

$$P(X_{V\setminus j}|do(x_j)) = \prod_{i\in V, i\neq j} P(x_i|x_{pa(i)}) = \frac{P(X_V)}{P(x_j|x_{pa(j)})} \tag{18}$$

I can now give some intuition (though not a full proof) for the back-door adjustment formula from section 7.6.3. Suppose we are interested in the causal effect of variable T on Y, for which the key ingredient is $P(Y|do(t)) = P(Y(t))$. Note that the parent set $pa(T)$ is always a valid back-door adjustment set (Exercise: why?). Let $X = pa(T)$ represent this adjustment set

and let W represent any other variables in the DAG. For simplicity, assume everything is discrete and let notation like $P(x)$ stand for $P(X = x)$.[37] Using equation 18, we have

$$\begin{aligned}
P(y, w, x|do(t)) &= \frac{P(y, w, x, t)}{P(t|x)} \\
P(y|do(t)) &= \sum_w \sum_x \frac{P(y, w, x, t)}{P(t|x)} \qquad \text{(marginalize out } w, x\text{)} \\
&= \sum_x \frac{P(y, x, t)}{P(t|x)} \\
&= \sum_x P(y|x, t)\frac{P(x, t)}{P(t|x)} \\
&= \sum_x P(y|x, t)\frac{P(x, t)}{\frac{P(t,x)}{p(x)}} \\
&= \sum_x P(y|x, t)P(x)
\end{aligned}$$

The last line is exactly the back-door adjustment formula! The proof for general adjustment sets is more complicated but follows a similar idea.

7.6.6 (Extra - not covered in Fall 2023): The do calculus

The do calculus consists of three core rules which allow us to link counterfactual distributions involving $do(X)$ to observed distributions for data from a given observational (non-intervention) DAG. These rules have been shown to be complete in the sense that if we cannot express our causal estimand $P(Y|do(x))$ in terms of observational quantities via the rules, then no such expression exists. We do not focus on these in the course but I will state and illustrate the first one.

First, define the following notation:

- $G_{\bar{X}}$: the DAG obtained by deleting all incoming arrows of X (i.e. the post $do(X)$ DAG)
- $(A \perp\!\!\!\perp B|C)_{G_{\bar{X}}}$: a conditional independence in graph $G_{\bar{X}}$

do-calculus rule 1:

$$\text{if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\bar{X}}}, \quad \text{then } P(y|do(x), z, w) = P(y|do(x), w)$$

Intuition: This rule concerns when we can insert or delete a variable we did not intervene on in the conditioning set. $G_{\bar{X}}$ reflects a world where we intervened $do(X)$ so Z is not an intervened-on variable. If, once we intervene on X, Y and Z are d-separated by X and W (aka, we've broken any links between them), then the rule says that we can drop Z from conditioning. Figure 22 gives a simple example where clearly, if we intervene on X, Z becomes irrelevant because it is only some cause of X in the non-intervention world. The rule does not apply to W, which maintains a direct relationship to Y.

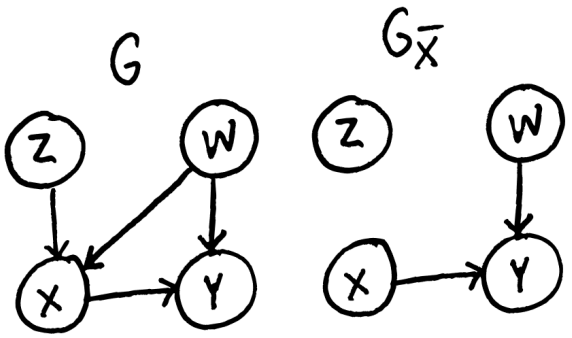


Figure 22: Example where do calculus rule 1 applies

The other two rules are a bit more complicated. Rule 2 describes when we can replace a $do(X)$ with just an X in a conditioning set while rule 3 describes when we can drop a $do(X)$ from a conditioning set entirely.

[37] All of this works for continuous distributions, where we would get pdfs $p(x)$.

7.6.7 DAG Resources

- http://dagitty.net/ - you can use this to check the conditional independences implied by a DAG
- Cunningham (2021) has a chapter on DAGs https://mixtape.scunning.com/03-directed_acyclical_graphs with some applied examples and discussion of the backdoor criterion

8 Module 8: Matching, Weighting, and Doubly-Robust Estimation

Big Picture: previously, given overlap and unconfoundness, we posed regression estimators which required us to model some functional form for Y. Matching and weighting still rely on overlap and unconfoundedess but provide a non-parametric alternative to modeling the outcome. Matching reflects the intuitive notion of comparing treated and control units which are similar on pre-treatment covariates. Weighting is a generalization that does not require us to sort units into discrete match and non-match groups. An important case of weighting is propensity score weighting, which involves a kind of parametric modelling but only of the probability of treatment rather than of outcome. Instead of choosing between regression-based approaches and weighting-based approaches, doubly-robust estimators combine them in a special way.

8.1 Matching

Matching approaches implement the intuitive notion that to see whether a treatment made a difference, we could compare pairs or groups of units which are similar, thereby controlling for other factors. It is **non-parametric** in the sense that it does not require specifying any model of the outcome and how the outcome varies with covariates X. Instead, for each treated unit i, we try to find $\mathcal{M}_i$, a set of indices of control units which approximately or exactly to match unit i in terms of *pre-treatment* covariates. Often with these approaches, we focus on the **ATT** (Average Effect on the Treated $\mathbb{E}(Y_i(1) - Y_i(0)|T_i = 1)$ so we only need $\mathcal{M}_i$ for each treated unit. We can then estimate the ATT using the actually observed treated outcomes $Y_i = Y_i(1)$ and a non-parametrically imputed $Y_i(0)$ based on the average of match set:

$$\hat{\tau}_{match} = \frac{1}{n_1}\sum_{i=1}^{n} T_i \left(Y_i - \frac{1}{|\mathcal{M}_i|}\sum_{i' \in \mathcal{M}_i} Y_{i'} \right) \tag{19}$$

This is again a kind of imputation estimator. It would be fine to flip this logic and estimate an ATC instead. There are also versions for other estimands.

8.1.1 Assumptions and Identification

We have not left the world of Module 7, in which **overlap** and **confoundedness** are the key identification assumptions. In module 7, the additional somewhat implicit assumption behind the regression estimators in Section 7.2 was that we have a correctly specified model $\mu_t(X_i) = \mathbb{E}(Y_i(t)|T_i = t, X_i)$. In this module, the implicit assumption is about the quality of the matching. The quantity $\mathbb{E}(Y_i(1)|T_i = 1)$ is immediately identified by consistency $\mathbb{E}(Y_i(1)|T_i = 1) = \mathbb{E}(Y_i|T_i = 1)$, but the quantity $\mathbb{E}(Y_i(0)|T_i = 1)$ is identified if the right term in equation 19 has expectation $E(Y_i(0)|T_i = 1)$. If, within match groups, treatment is randomized, then we have a mini-experiment within match group and the left term in the above estimator will be unbiased for $E(Y_i(0))$. If there are unobserved confounders making the $Y_j(0)$'s for $j \in \mathcal{M}_i$ systematically different from $Y_i(0)$, that is a problem. As discussed in detail in section 8.1.7, even if unconfoundedness does hold at the exact $X_i = x$ level $(Y_i(0), Y_i(1) \perp T_i|X_i = x)$, if matching is imperfect, unconfoundedness may not hold perfectly within a match group and some bias can be introduced. As discussed in Section 8.1.3, when estimating the ATT, lack of overlap coming from control units in regions of covariate space whether there are no treatment units can be ok and matching can help us deal with this.

8.1.2 Matching and Covariate Balance

The key goal of matching is to achieve **balance**, which refers to the treatment and control groups having the same distributions over X. Checking balance is a key way that people evaluate whether a matching method has worked well. Logically, if each treated unit were matched to a control unit with exactly the same X_i, then balance follows. You might be worried that technically, good balance does not imply good individual-level matching. Imagine, for example, one-to-one matching n_1 treated units exactly to n_1 control units and then randomly permuting the controls. Then the distributions would stay the same but the individual matches could be terrible! However, actually, for the ATT estimator in Equation 19, this could make little or no difference. In particular, $|M_i| = c$ for some constant c, we can then write

$$\hat{\tau}_{match} = \frac{1}{n_1}\sum_{i=1}^{n} T_i Y_i - \frac{1}{n_1 c}\sum_{i:T_i=1}\sum_{i' \in \mathcal{M}_i} Y_{i'}$$

and the right term would be the same even if the individual matches were decoupled! For this reason and because of the ideas discussed in Section 8.1.3, it makes sense to think of balance as a primary goal.

8.1.3 Matching as non-parametric pre-processing

One way to think of matching is as a kind of non-parametric preprocessing that can reduce sensitivity to outcome model mis-specification and prevent extrapolation by removing outlying and potentially distorting control units. In Figure 23 (left), the global regression of Y_i on X_i is sensitive to whether we include a quadratic X_i^2 term or not, but it is mainly control units far from any treated units that are creating the quadratic structure. The right side of the figure shows that if instead we fit the two regression models only in a window of control units *near* the treated units in X, then the model choice does not make much difference. The simplest case matching estimator above is not actually fitting a regression like this, but one application of matching algorithms can be to remove any control units that do not get matched and then run the regression. One could also run a local regression within each match group (see Section 8.1.7). The overall intuition is that by focusing on local comparisons, we hope to avoid distortions from 'apples-to-oranges' comparisons. This is formalized in the following section.

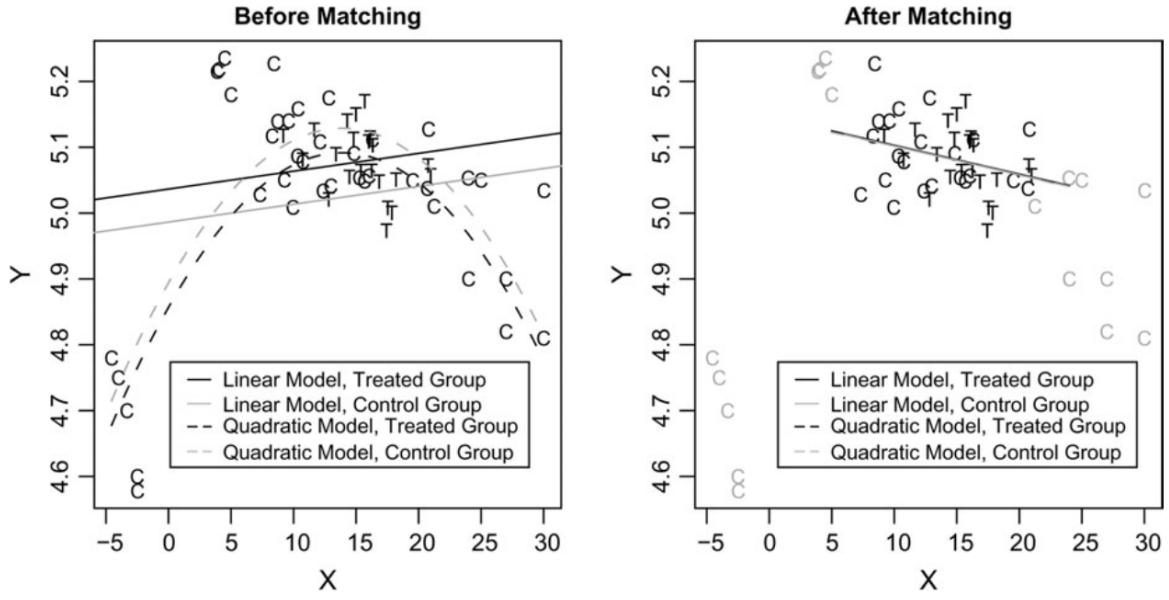


Figure 23: Figure copied from Imai lecture Slide 3

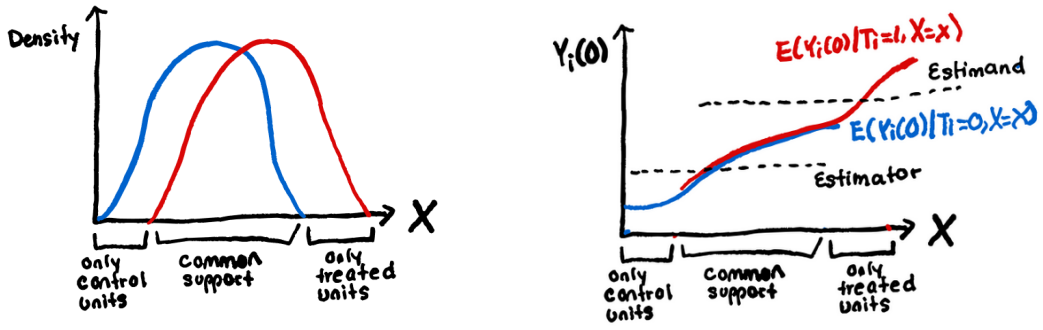


Figure 24: **Left:** Illustration of common support of treatment and control group for case of univariate X. **Right** illustration of a scenario that would create bias due to lack of common support. The highest $Y_i(0)$ are only in the treated group and the lowest $Y_i(0)$'s are only in the control group and vice versa. This pulls up and down the averages – the black dotted lines represent possible average values. Their difference represents bias.

8.1.4 Bias: How can matching help?

When estimating the ATT, the key unknown to identify is $\gamma := \mathbb{E}(Y_i(0)|T_i = 1)$. Suppose we do not match and simply use the control group sample mean, which is an unbiased estimator of $\mathbb{E}(Y_i|T_i = 0)$. We will explore the bias that could arise from this and where matching can help. First, define the following notation:

- $F_{X_i|T_i=t}(x)$ the CDF of X within the $T_i = t$ group
- S_1 the support of X for the treatment group – i.e. the values such that $P(X_i = x|T_i = 1) > 0$ or pdf $p(x|T_i = 1) > 0$.
- S_0 the support of X for the control group.
- $S = S_1 \cap S_0$, the common support for treatment and control group (see Figure 24)

The bias expression is then[38]

$$\begin{aligned} bias &= \mathbb{E}(Y_i|T_i = 0) - \mathbb{E}(Y_i(0)|T_i = 1) \\ &= \mathbb{E}(Y_i(0)|T_i = 0) - \mathbb{E}(Y_i(0)|T_i = 1) && \text{(by consistency)} \\ &= \mathbb{E}(\mathbb{E}(Y_i(0)|T_i = 0, X_i)) - \mathbb{E}(\mathbb{E}(Y_i(0)|T_i = 1, X_i)) && \text{(Law of Total Expectation)} \end{aligned}$$

As on Slide 4, We can decompose this bias into three parts $bias = A + B + C$ where[39]

1. **Part A:** involves the part of each expectation that is not coming from the common support and captures bias from lack of common support if there is any. A lack of common support reflects a violation of the **overlap condition** that $1 > P(T_i = 1|X_i = x) > 0$ for all x (see Section 7.1). It indicates there are x where $P(T_i = 1|X_i = x) = 0$ or 1.

$$A = \int_{S_0 \setminus S} \mathbb{E}(Y_i(0)|T_i = 0, X_i = x) dF_{X_i|T_i=0}(x) - \int_{S_1 \setminus S} \mathbb{E}(Y_i(0)|T_i = 1, X_i = x) dF_{X_i|T_i=1}(x)$$

   ✓Matching deals with this by restricting to comparisons in the area of commons support. Caution: if $S_1 \neq \emptyset$, this does change our ATT estimand. If the only lack of common support comes from the control group as in Figure 23, then we might remove these controls without biasing the ATT estimator. See 24 (right) for an illustration.

2. **Part B:**[40] reflects bias due to **imbalance** in the X distribution within the area of common support. Even if $\mathbb{E}(Y_i(0)|T_i = 0, X_i = x) = \mathbb{E}(Y_i(0)|T_i = 1, X_i = x)$ within the area of common support (Figure 24-right) an imbalance in this region (Figure 24-left) could make $\mathbb{E}(\mathbb{E}(Y_i(0)|T_i = 0, X_i)) \neq \mathbb{E}(\mathbb{E}(Y_i(0)|T_i = 1, X_i))$

$$B = \int_S \mathbb{E}(Y_i(0)|T_i = 0, X_i = x)[dF_{X_i|T_i=0}(x) - dF_{X_i|T_i=1}(x)]$$

   ✓Matching (and weighting below) both aim to improve balance.

3. **Part C:** reflects bias due to unobservable (confounding) variables that mean that even within the area of common support, the $Y_i(0)$'s are systematically different within the treated group and control group.

$$C = -\int_S \left(\mathbb{E}(Y_i(0)|T_i = 1, X_i = x) - \mathbb{E}(Y_i(0)|T_i = 0, X_i = x)\right) dF_{X_i|T_i=1}(x)$$

   ×Matching and weighting do not fix violations of unconfoundedness! Balance on observables need not mean balance on unobservables.

[38] Slide 4 has $\mathbb{E}(Y_i(0)|T_i = 1) - \mathbb{E}(Y_i|T_i = 0)$, which would generally be called the negative bias. I flip this to the standard version so a few signs are flipped relative to the slides.

[39] **Notation note:** the $dF(x)$ notation in the integrals here is just a more general notation that allows this integral to apply to either discrete X (where we'd usually write a sum) or continuous X. If X is continuous with pdf $f(x)$, then $dF(x) = f(x)dx$ and we have $\mathbb{E}(X) = \int x dF(x) = \int x f(x) dx$. If you are more used to the latter form, do not worry about the former. If you'd like to learn more about the former, you can look up Lebesgue integrals – but these are not needed for this course!

[40] This comes from doing a +0 trick by subtracting a $\mathbb{E}(Y_i(0)|T_i = 0, X_i = 0) dF_{X_i|T_i=1}(x_i)$ in part B and adding it in Part C.

8.1.5 Some Matching Methods

The following is a non-exhaustive list of some methods involved in forming matches:

1. **Exact Matching:** Here for each treated unit i, its match group is the control units which match on covariates X_i exactly: $\mathcal{M}_i = \{j : T_j = 0, X_j = M_j\}$. This approach may be feasible for a small number of discrete covariates X_i and if feasible, is attractive since it achieves balance perfectly and fully accounts for information in X with no need to consider any further regressions on X_i as we might want to do in Figure 23. However, for continuous covariates or many discrete ones, exact matching quickly becomes infeasible and too many treated units may end up with no matches.
2. **Coarsened Exact Matching (CEM)**: This follows the same logic as the previous only with some discretizing of continuous variables and collapsing of some discrete categories to make the matching task feasible. There is still the risk that some treated units end up with no match. There is a bias-variance trade-off here where coarser categories make matching more possible and lower the variance but also introduce more bias. We might want to consider the strategy discussed in section 8.1.7.
3. **Distance Metric**: Let X_i be the vector of covariates for treated unit i and X_j be the same for control unit j. If we define some distance metric $d(X_i, X_j)$ such as Mahalanobis Distance, then we might do matching to reduce the distance between X_i and its matches. This is a kind of dimension reduction where we try to reduce comparisons in high-dimensional space to some single similarity/distance metric. This approach includes many different options, including different choices of distance metrics and what matching algorithm to apply to the distances (e.g., nearest neighbor). There is also a choice whether to do one-to-one or one-to-many matching and whether to allow different treated units to match to the same control unit (with/without replacement). One might also set a threshold and refuse to match units with too large distances.
4. **Propensity Score Matching**: one possible 'distance metric' we could use for matching is difference in units' (estimated) propensity score. The propensity score is a unit's probability of receiving treatment $\pi_i(X_i) := Pr(T_i = 1|X_i)$. We will discuss propensity scores more in the context of weighting, but as a matching method, they again form a kind of dimension reduction and have some nice balancing properties that make it desirable to compare units with similar $\pi(X_i)$'s. See Section 8.2.1. Note, however, that some warn against using this matching metric (see King and Nielsen (2019) for a discussion, including a comparison to Mahalanobis Distance).
5. **Optimal Matching Algorithms**: one can also approach matching from an optimization perspective. Various algorithms exist for determining matches, subject to some constraints and sometimes conditional on some choice of distance metric. One approach formulates matching as a minimum cost flow problem where the goal is to minimize the sum of pairwise distances between treated units (Rosenbaum, 1989). Another, called cardinality matching, sets some thresholds (tolerances) for the covariate balance in each covariate and then tries to maximize the number of matched units subject to that tolerance and to a one-to-one matching constraint. The goal is to match as many units as possible while also meeting the thresholds (Niknam and Zubizarreta, 2022; Visconti and Zubizarreta, 2018).

There are different guidelines for how to pick a matching method. One sensible approach is to mix the approaches above. For example, we might first decide on a few variables for which to require exact matches (e.g., decide that we will only match tutored children to untutored children in the same SES level) and then use some context-appropriate distance metric to determine a more specific match group based on additional covariates which are less fundamental.

Dropping Unmatched Units: Matching methods generally can result in some units having no matches and therefore being dropped. If these units are control units, there can sometimes be some information loss, but it will not bias the ATT or change the estimand. However, if treated units are dropped, this biases estimates of our original ATT or, put another way, changes our estimand to not include those types of units.

8.1.6 Checking Covariate Balance

Regardless of which matching method you choose, it is important to evaluate the quality of the resulting match by checking how close the treated unit and its match group are on pre-treatment covariates and in particular, checking covariate balance. Ideally, we'd like to look at the joint distribution of all covariates X between the treatment and control groups, but this becomes difficult for even a moderate number of covariates. In practice, people tend to check **lower-dimensional summaries** such as means and variances of marginal distributions. Common options include

1. **Standardized mean difference**: Suppose we have p covariates so that each unit i has X_{ij} for $j = 1, ..., p$. For each covariate, we define:

$$\text{smd}_{j,treat} = \frac{\frac{1}{n_1}\sum_{i=1}^{n} T_i \left(X_{ij} - \bar{X}_{ij}\right)}{S} \tag{20}$$

   where

   - $\bar{X}_{ij} = \frac{1}{\mathcal{M}_i}\sum_{i' \in \mathcal{M}_i} X_{i'j}$ is the mean of covariate j in group of matched controls for unit i $\mathcal{M}_i$
   - $S = \sqrt{\frac{1}{n_1 - 1}\sum_{i=1}^{n} T_i (X_{ij} - \bar{X}_{j1})^2}$ is the standard deviation of covariate j within the treatment group
   - $\bar{X}_{j1} = \frac{1}{n_1}\sum_{i=1}^{n} T_i X_{ij}$ is the mean of covariate j in the treatment group

   In the numerator of this quantity, we are looking at the average difference between unit i's X_{ij} and the average match group X_{ij} value. We want these to be small. But how small is small? S standardizes this value by the baseline noise in X_{ij} in the treatment group. We focus on the treatment group here because we care about the ATT. As in Figure 23, it might not be a problem for matching if the control group has higher variance. As an example, imagine one covariate is age. If on average, treated units differ from their matches by 5 years in a treatment group of only 60-70 year olds, that seems like bad matching. But if the treatment group has 18-90 year olds, that might be more reasonable. A rough rule of thumb sometimes used in practice is that smd values on the order of .1 are good, and .2 are ok – but this can vary by application. In `R`, see the `Cobalt` package (`bal.tab` function) for help implementing this.

2. **Empirical CDF comparison**: while smd focuses on mean and variance, a way to look at the whole distribution is to calculate empirical CDFs for each covariate within the post-matching treatment and control groups

$$F_{jt}(x) = \frac{1}{n_t}\sum_{i=1}^{n} I(T_i = t) I(X_{ij} \leq x)$$

   If, via matching, we have thrown out some units with no matches, we may thereby have made the distributions more similar. **Caution:** this is not appropriate for all matching. For example, if we have done matching with replacement, where multiple treated units may be matched to the same control, the ECDFs will not reflect this.

   Still, in settings where appropriate, **visualization** of the histogram or ECDF can be an important way to get some sense of how well balance holds.

3. **Testing:** it might seem like a good idea to formalize the task of deciding whether there is a difference in distributions by running a hypothesis test.

   Do not use: Balance Tests: these are tests such as the t-test for a difference in means which define a null of there being no difference and reject if the observed difference is large. However, this approach can be misleading. As usual in a hypothesis test, *failing to reject the null* does not mean the null of covariate balance is correct. For one thing, there may be distributional differences the test does not capture that would become obvious to you if you plotted the histogram or looked at some statistical summaries. Another issue is that matching may reduce your sample size and thereby reduce your statistical power to detect a difference – you may end up with poor balance but large enough standard errors that you do not reject your null. In other words, balance tests create a reward for dropping many observations and thereby hurting your power so that you do not reject the null.

   Do use: There is a better approach called **equivalence testing** that looks if we can reject a null hypothesis that there *is* a difference of at least Δ (representing e.g. a difference in means). This shifts the burden of proof – is the similarity so high that it is implausible for it to be generated under a situation where truly, the distributions are at least Δ apart? One can also form an interval reflecting the range of differences that we can rule out (See Wellek (2010) for the theory, Dunning and Nilekani (2013) for an example application).

8.1.7 Remaining Bias: Correcting for Imperfect Covariate Balance (Slide 9)

Matching usually is not perfect, and there can be remaining covariate imbalance that creates bias. Consider the estimand $\mathbb{E}(Y_i(0)|T_i = 1) = \mathbb{E}(\mathbb{E}(Y_i(0)|T_i = 1, X_i))$ and in particular, define:

$$\mu_{0t}(X_i) := \mathbb{E}(Y_i(0)|T_i = t, X_i)$$

In matching, we are trying to estimate $\mu_{01}(X_i)$ using values of $\mu_{00}(X_j)$ for j in the match group $\mathcal{M}_i$. Let $\mathcal{X}_{\mathcal{M}_i}$ refer to the set of $X_{i'}$ covariate values for match group $\mathcal{M}_i$. Then the (conditional) bias of estimating $\mu_{01}(X_i)$ using $\mu_{00}(X_j)$ for $i' \in \mathcal{X}_{\mathcal{M}_i}$ is:[41]

$$\text{Bias}(X_i, \mathcal{X}_{\mathcal{M}_i}) = \mathbb{E}\left(\frac{1}{|\mathcal{M}_i|}\sum_{i' \in \mathcal{M}_i} Y_{i'} \mid \mathcal{X}_{M_i}\right) - \mu_{01}(X_i) = \left(\frac{1}{|\mathcal{M}_i|}\sum_{i' \in \mathcal{M}_i} \mu_{00}(X_{i'})\right) - \mu_{01}(X_i)$$

Note that under **unconfoundedness**, $\mu_{01}(X_i) = \mu_{00}(X_i)$ for all X_i.[42] Assume this holds. If $X_{i'} = X_i$ for all i', we then we have no bias. However, if there are still some differences between $X_{i'}$ and X_i, this could create bias. One approach to **bias correction** is to re-introduce some regression within the match group. We use this to estimate the bias and then add it back into our estimator.

Formally, let A be the set of control units that are matched to at least one treated unit $A = \{i' : \exists i : i' \in M_i\}$. Consider regressing $Y_{i'}$ on $X_{i'}$ just in A: $Y_{i'} = X_{i'}^T\beta + \epsilon$.[43] The predicted value at X_i for this regression is then $\hat{Y}_i = X_i^T\hat{\beta}$ and this may differ from the fitted value $\hat{Y}_{i'} = X_{i'}^T\hat{\beta}$ as illustrated in Figure 25. We can treat this difference as an estimate of the bias from using units with $X_{i'}$ to estimate the expected value of $Y_i(0)$ for units with X_i. That is: define

$$\hat{B}(X_i, X_{i'}) = \hat{\beta}^T(X_i - X_{i'})$$

We can then define the estimator

$$\begin{aligned}\hat{\mu}_{00}(X_i) &= \frac{1}{|\mathcal{M}_i|}\sum_{i' \in \mathcal{M}_i}\left(Y_{i'} + \hat{\beta}^T(X_i - X_{i'})\right) \\ &= \frac{1}{|\mathcal{M}_i|}\sum_{i' \in \mathcal{M}_i}\left(Y_{i'} + \hat{B}(X_i, X_{i'})\right)\end{aligned}$$

If $\hat{B}(X_i, X_{i'})$ is unbiased for $\mu_{00}(X_i) - \mu_{00}(X_{i'})$, then this removes the bias from imperfect balancing. See Abadie and Imbens (2011) for more.

[41] Technicalities: $\mu_{0t}(X_i)$ can be considered as super-population quantities. Hence even if in our sample, there is only one unit with $X_i = x$, we can think of a population average at $X_i = x$. In the calculations in this section, the match groups are treated as fixed rather than as random variables. We are conditioning on X_i, treatment status, and the matching.

[42] This is why Slide 9 simply calls it $\mu_0(X_i)$.

[43] Note: if we have many matched control units for each treated units, we could also consider an even more local regression where we regress $Y_{i'}$ on $X_{i'}$ separately within each M_i group and calculate a $\hat{\beta}_i$ specific to treated unit i, but often, we will not have enough data for this to be a reasonable prospect.

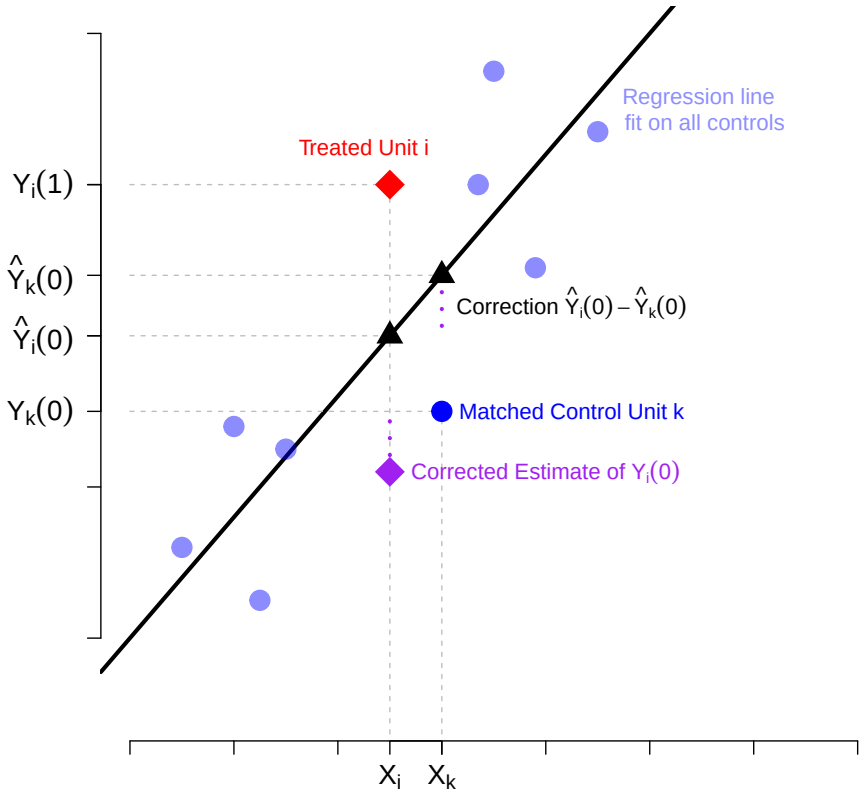


Figure 25: Illustration of bias correction idea. Suppose we have a unit i matched to unit k because its X_k is the closest available to X_i. The fact that $X_i \neq X_k$ may create error in our estimate of $Y_i(0)$ if there is a strong correlation between X and Y. As the diagram illustrates, instead of using $Y_k = Y_k(0)$ (height of the dark blue circle) as an estimate of $Y_i(0)$, we use $Y_k - (\hat{Y}_i - \hat{Y}_k)$ (height of the purple diamond) where this correction term comes from the regression model.

Connection: recall in Module 6, Section 6.6 (Figure 13), we considered what would happen if treatment were randomized within a window around the RD cut-off point and the issue was that if there was a *slope*, assuming as-if-randomness and taking a difference in means in the window could distort conclusions. Here, there is something similar going on. In Figure 8.1.7, X and Y have a positive slope. This means that even controls close to X_i may have systematically higher or lower expected $Y_i(0)$ values. The standard matching estimator is taking a mean in the matching window, which would be fine if there were a flat slope. The bias correction is accounting for the slope.

> "Matching can be used in conjunction with model-based adjustment. The role of the matching is to make the treatment and control group sufficiently similar so that model-based adjustment can be done locally for remaining imbalance of X" (Kosuke Imai, Stat 286 Video Lecture 8-2)

8.1.8 Matching as Weighting Estimator

The matching estimator of the ATT in Equation 19 can be algebraically manipulated to the following form:

$$\begin{aligned}\hat{\tau}_{match} &= \frac{1}{n_1}\sum_{i=1}^{n} T_i\left(Y_i - \frac{1}{|\mathcal{M}_i|}\sum_{i'\in\mathcal{M}_i} Y_{i'}\right)\\ &= \frac{1}{n_1}\sum_{i=1:T_i=1} Y_i - \frac{1}{n_0}\sum_{i:T_i=0} w_i Y_i\end{aligned}$$

where

$$w_i = \frac{n_0}{n_1}\sum_{i':T_{i'}=1}\frac{I(i\in\mathcal{M}_{i'})}{|\mathcal{M}_{i'}|}$$

the main part of the weight comes from noting that each $Y_{i'}$ will be included and multiplied by a $\frac{1}{|\mathcal{M}_i|}$ term every time that it is present in the $\mathcal{M}_{i'}$ for some treated i'. We simply add these up across the treated units to get the weight. The weights therefore reflect how often each control unit is matched to a treatment unit and the size of the matching groups. For example, if we have a one-to-one matching scheme in which each treated unit is matched to one control unit without replacement, the weights are all 0 or 1. If each treated unit is matched to 5 control units without replacement, then the weights are all $\frac{1}{5}$. With replacement and variable match group sizes, things get trickier to calculate. Notice that the w_i sum to n_0 so the 'total' weight of the control group is not changed. Implicitly, for the treated group, we have $w_i = 1$.

8.1.9 Variance

Variance can be an issue with matching (and weighting) estimators and doing variance calculations for matching and weighting estimators can get tricky because of the fact that the weights and match groups are also random variables. and We provide only a brief look at this area, which is an active topic of research. For the matching case, consider the estimator as a weighting estimator

$$\hat{\tau}_{match} = \frac{1}{n_1}\sum_{i=1}^{n} T_i Y_i - \frac{1}{n_0}\sum_{i=1}^{n}(1-T_i)W_i Y_i$$

where W_i is defined as in Section 8.1.8. Assume unconfoundedness and let $\mu_t(X_i) = \mathbb{E}(Y_i(t)|T_i = 1, X_i) = \mathbb{E}(Y_i(t)|T_i = 0, X_i)$. If we consider the X_i and the matching weights as random, then calculating the variance of this gets complicated. If we consider the matching as fixed and condition on X_i and treatment status T_i, then we can at least do some calculations for the *conditional* variance of $\hat{\tau}_{match}$ coming from the randomness in Y_i. From the random-Y, fixed-X perspective, we can think of $\hat{\tau}$ as representing an estimate of the average **Conditional Average Treatment Effect on the Treated** (CATT) averaged over the particular values of X_i among the treatment group

$$CATT = \frac{1}{n_1}\sum_{i=1}^{n} T_i(\mu_1(X_i) - \mu_0(X_i))$$

The estimation error of $\hat{\tau}_{match}$ relative to this CATT is then

$$\begin{aligned}\hat{\tau}_{match} - CATT &= \left(\frac{1}{n_1}\sum_{i=1}^{n} T_i Y_i(1) - \frac{1}{n_0}\sum_{i=1}^{n}(1-T_i)W_i Y_i(0)\right) - \left(\frac{1}{n_1}\sum_{i=1}^{n} T_i(\mu_1(X_i)-\mu_0(X_i))\right)\\ &= \left(\frac{1}{n_1}\sum_{i=1}^{n} T_i Y_i(1) - \frac{1}{n_0}\sum_{i=1}^{n}(1-T_i)W_i Y_i(0)\right) - \left(\frac{1}{n_1}\sum_{i=1}^{n} T_i(\mu_1(X_i)-\mu_0(X_i))\right) \pm \frac{1}{n_0}\sum_{i=1}^{n}(1-T_i)W_i\mu_0(X_i)\\ &= \left(\frac{1}{n_1}\sum_{i=1}^{n} T_i\mu_0(X_i) - \frac{1}{n_0}\sum_{i=1}^{n}(1-T_i)W_i\mu_0(X_i)\right) + \left(\frac{1}{n_1}\sum_{i=1}^{n} T_i(Y_i(1)-\mu_1(X_i)) - \frac{1}{n_0}\sum_{i=1}^{n}(1-T_i)W_i(Y_i(0)-\mu_0(X_i))\right)\end{aligned}$$

Because we are only subtracting a constant, the conditional variance of $\hat{\tau}_{match}$ is equivalent to conditional variance of this. The form above is useful because the first term should, under good matching and a large enough sample size, be ≈ 0 so we will ignore it. Treating X_i and T_i as fixed, we can then calculate the variance of the second term, which is like a sum of residuals. Using the assumption that we have i.i.d. observations and hence no cross term below, we get

$$
\begin{aligned}
\mathbb{V}(Y_i|T,X) &\approx \frac{1}{n_1^2}\sum_{i=1}^{n} T_i \mathbb{V}(Y_i(1)|T,X) + \frac{1}{n_0^2}\sum_{i=1}^{n}(1-T_i)W_i^2\mathbb{V}(Y_i(0)|T,X) \\
&= \frac{1}{n_1^2}\sum_{i=1}^{n} T_i \mathbb{V}(Y_i|T,X) + \frac{1}{n_0^2}\sum_{i=1}^{n}(1-T_i)W_i^2\mathbb{V}(Y_i|T,X) \qquad \text{(consistency)} \\
&= \left(\sum_{i=1}^{n}\frac{T_i}{n_1^2} + \frac{1}{n_0^2}(1-T_i)W_i^2\right)\mathbb{V}(Y_i|T,X) \\
&= \left(\sum_{i=1}^{n}\frac{T_i}{n_1} + \frac{1}{n_0}(1-T_i)W_i\right)^2\mathbb{V}(Y_i|T,X)
\end{aligned}
$$

The $\mathbb{V}(Y_i|T,X)$ term is itself estimable via

1. Matching methods that work within the treatment and within the control group as described in (Imbens and Rubin, 2015, ch. 19) – we need this because we need multiple observations for a given X to estimate the conditional variance and, especially if X is continuous, do not have these immediately.
2. Heteroskedasticity-robust standard errors from regression

8.2 Propensity Score Weighting

Weighting methods have many of the same intuitions and motivations as matching methods and can be viewed as a generalization of matching. In matching methods, we link each treated individual unit i to a set of control units. A major goal here is covariate balance. However, drawbacks of matching methods include that they may end up throwing away some data (e.g., a control unit which is not the closest match for any treated unit) and they are restricted to discrete decisions (match or no match) that may not be able to optimally balance covariates. Weighting is more flexible. As discussed below, weighting also comes with a risk of high variance if some weights are extreme.

8.2.1 Propensity Scores

The **propensity score** of unit i is defined as $\pi(X_i) = \Pr(T_i = 1|X_i)$. In the experimental contexts of Modules 1-4, we knew these exactly (for CRD, we had $\Pr(T_i = 1) = \frac{n_1}{n}$) but in observational studies, we do not know the exact assignment mechanism. This means $\pi_i(X_i)$ must be estimated, for example via a logistic regression or some more flexible ML classifier. The *theoretical* propensity score $\pi(X_i)$ has attractive properties:

1. **Balancing Property**

$$T_i \perp X_i|\pi(X_i)$$

This says that once we condition on propensity score, covariates no longer predict anything about treatment status. Note that this says nothing about any unobserved variables U_i.

Proof. This is mathematically true without any assumptions. We want to show that $\Pr(T_i = 1|\pi(X_i), X_i) = \Pr(T_i = 1|\pi(X_i))$. Because conditioning on $\pi(X_i)$ if we already know X_i is redundant,

$$\Pr(T_i = 1|\pi(X_i), X_i) = \Pr(T_i = 1|X_i) = \pi(X_i)$$

Applying the Law of Total Expectation gives

$$\begin{aligned}\Pr(T_i = 1|\pi(X_i)) &= \mathbb{E}(T_i|\pi(X_i)) \\ &= \mathbb{E}(\mathbb{E}(T_i|\pi(X_i), X_i)|\pi(X_i)) \\ &= \mathbb{E}(\mathbb{E}(T_i|X_i)|\pi(X_i)) \\ &= \mathbb{E}(\pi(X_i)|\pi(X_i)) = \pi(X_i)\end{aligned}$$

□

2. **Dimension reduction / relation to confoundedness**. If unconfoundedness $(Y_i(1), Y_i(0) \perp T_i|X_i)$ and overlap $(\pi(x) \in (0, 1)\ \forall x)$ hold, then

$$Y_i(1), Y_i(0) \perp T_i|\pi(X_i)$$

This says that once we condition on propensity score, we essentially have a mini-randomized experiment among those with the same score with treatment independent of potential outcomes.

Proof. The goal is to show $\Pr(T_i = 1|Y_i(1), Y_i(0), \pi(X_i)) = \Pr(T_i = 1|\pi(X_i))$.

$$\begin{aligned}\Pr(T_i = 1|Y_i(1), Y_i(0), \pi(X_i)) &= \mathbb{E}(T_i|Y_i(1), Y_i(0), \pi(X_i)) \\ &= \mathbb{E}(\mathbb{E}(T_i|Y_i(1), Y_i(0), \pi(X_i), X_i)|Y_i(1), Y_i(0), \pi(X_i)) \quad \text{(Law of total expectation)} \\ &= \mathbb{E}(\mathbb{E}(T_i|Y_i(1), Y_i(0), X_i)|Y_i(1), Y_i(0), \pi(X_i)) \quad \text{(redundancy)} \\ &= \mathbb{E}(\mathbb{E}(T_i|X_i)|Y_i(1), Y_i(0), \pi(X_i)) \quad \text{(unconfoundedness)} \\ &= \mathbb{E}(\pi(X_i)|Y_i(1), Y_i(0), \pi(X_i)) \\ &= \pi(X_i) \\ &= \Pr(T_i = 1|X_i) \\ &= \Pr(T_i = 1|X_i, \pi(X_i)) \quad \text{(redundancy)} \\ &= \Pr(T_i = 1|\pi(X_i)) \quad \text{(balancing)}\end{aligned}$$

□

In theory, these properties justify using $\pi(X_i)$ as a much lower dimensional co-variate to condition on or use as a metric for matching. Conditional on units with the same propensity score, we have a 'mini-randomized experiment'! The BIG caveat to these properties is that the *estimated* propensity score we have to use in practice may not satisfy these properties! If the propensity scores are estimated badly, we have no guarantee they provide balance. Given estimated propensity scores, we can use the balancing property as a diagnostic. If, conditional on the estimated propensity scores, we have evidence that T and X are not independent, that indicates a problem with the propensity scores.

8.2.2 The IPW Estimator

The **Inverse Probability Weighting** (IPW) estimator of the ATE uses estimated propensity scores to down-weight treated observations with a higher probability of receiving treatment and upweight those with a low probability of treatment – and vice versa for the control group. The estimator is exactly the **Horvitz Thompson Estimator** developed in the context of survey sampling by Horvitz and Thompson (1952).[44] In the case of the ATT and ATC, there are slight tweaks, but the idea is similar. Table 5 gives the different estimators.

Estimand	Estimator
ATE	$\frac{1}{n}\sum_{i=1}^{n}\frac{T_iY_i}{\hat{\pi}(X_i)} - \frac{(1-T_i)Y_i}{1-\hat{\pi}(X_i)}$
ATT	$\frac{1}{n_1}\sum_{i=1}^{n} T_iY_i - \frac{\hat{\pi}(X_i)(1-T_i)Y_i}{1-\hat{\pi}(X_i)}$
ATC	$\frac{1}{n_0}\sum_{i=1}^{n}\frac{(1-\hat{\pi}(X_i))T_iY_i}{\hat{\pi}(X_i)} - (1-T_i)Y_i$

Table 5: IPW Estimators for common estimands. See Fan Li and Zaslavsky (2018) for more details.

Suppose we replace the $\hat{\pi}(X_i)$ in the IPW estimator by their true values. Then is sometimes called the **oracle** IPW estimator for the ATE and it is unbiased by the following calculation,[45]

$$\begin{aligned}
\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{T_iY_i}{\pi(X_i)}\right) &= \mathbb{E}\left(\mathbb{E}\left(\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{T_iY_i(1)}{\pi(X_i)}|\mathcal{O}_n, X_i\right)|X_i\right)\right) \qquad \text{(Law of Tot. Exp. x2)}\\
&= \mathbb{E}\left(\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i(1)}{\pi(X_i)}\mathbb{E}(T_i|\mathcal{O}_n, X_i)\right)|X_i\right)\\
&= \mathbb{E}\left(\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i(1)}{\pi(X_i)}\pi(X_i)|X_i\right)\right) \qquad \text{(Unconfoundedness)}\\
&= \mathbb{E}\left(\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}Y_i(1)|X_i\right)\right)\\
&= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}Y_i(1)\right)\\
&= \mathbb{E}(Y_i(1))
\end{aligned}$$

See Module 8 Review Questions for the proof that the oracle ATT estimator is unbiased. In practice, we replace $\pi(X_i)$ with $\hat{\pi}(X_i)$ and if these estimates are poor, the estimator can be biased. We hope that if $\hat{\pi}(X_i)$ are at least consistent, then the above estimator is consistent. For high-dimensional X_i, estimating propensity scores can be hard. What to do in that situation is an active area of research (See e.g., Wang and Shah (2020)).

8.2.3 Normalized IPW Weights

Imagine that we have data on math scores, tutoring, and SES and unconfoundedness conditioning on SES is plausible. Suppose, however, that children from low SES backgrounds were unlikely to get tutoring. In that case, the few observations we *do* have for children from low SES backgrounds will be upweighted in the IPW estimator with their values counting as 'representatives' for the rare group. Of course, that also means that our estimator becomes sensitive to the particular outcomes for those children. The more skewed the weights $\pi(X_i)$ are, the higher the variance of the IPW estimator. In the extreme, if we have some $\pi(X_i)$ close to 0 or 1, this can lead to unstable behavior with a few units' values dominating the estimator. This is the motivation for the **Hajek Estimator**, which normalizes the weights so they sum to 1 for the treatment group and the control group.

$$\widehat{ATE} = \frac{\sum_{i=1}^{n}\frac{T_iY_i}{\hat{\pi}(X_i)}}{\sum_{i=1}^{n}\frac{T_i}{\hat{\pi}(X_i)}} - \frac{\sum_{i=1}^{n}\frac{(1-T_i)Y_i}{1-\hat{\pi}(X_i)}}{\sum_{i=1}^{n}\frac{1-T_i}{1-\hat{\pi}(X_i)}}$$

[44] In survey sampling, the intuitions are the same except instead of probability of treatment weights, we have probability of being selected into the sample.

[45] It is also valid to leave out the O_n notation and do law of total expectation with X_i only and factor into $E(T_i|X_i)$ and $E(Y_i(1)|X_i)$ immediately.

where for the left term, we are multiplying $T_i Y_i$ by the weight

$$w_i = \frac{1}{\hat{\pi}(X_i) \sum_{i=1}^n \frac{T_i}{\pi(X_i)}}$$

with

$$\sum_{i=1}^n T_i w_i = \frac{\sum_{i=1}^n \frac{T_i}{\hat{\pi}(X_i)}}{\sum_{i=1}^n \frac{T_i}{\hat{\pi}(X_i)}} = 1$$

and similarly for the control group weights. This leads to some bias, but potentially has a variance benefit. Other approaches to extreme weights include trimming them with some threshold.

Another nice property of the Normalized IPW estimator is that it can be shown to be equivalent to **weighted least squares** with the (normalized) weights above. In the case of the ATT, the weights for the treated units are implicitly $\frac{1}{n_1}$ (these clearly sum to 1 for the treated units) and in the case of the ATC, the weights for the control units are implicitly $\frac{1}{n_0}$.

In practice, it is usually a good idea to normalize your weights!

8.2.4 IPW Variance

Quick Review: Method of Moments Estimators (MoM): MoM estimators are created by the following steps:

1. Specify some **moment condition**. Often these come from calculating expectations which are functions of the parameter of interest or specifying some condition you want a certain expectation to meet.
2. Solve the equation(s) for the parameter(s) of interest
3. Replace true moments with sample moments to form estimator

For example, if $X_1, ..., X_n \overset{iid}{\sim} exponential(\theta)$ then one can show that $E_\theta(X) = \frac{1}{\theta}$. Solving this for θ yields $\theta = \frac{1}{E_\theta(X)}$. The method of moments estimator then results from plugging in $\bar{X}_n$ for $E_\theta(X)$ so we get $\hat{\theta} = \frac{1}{\bar{X}}$. The nice thing about MoM estimators is that IF you can specify an estimator as a MoM estimator, then under some regularity conditions, various standard asymptotic results such as asymptotic normality and asymptotic variance calculations apply.[46]

IPW Estimator as Methods of Moments Estimator

Suppose we have some model $\pi_\theta(X_i)$ of the propensity scores with unknown parameter θ. For example, this could be a logistic regression where the model is

$$\pi_\theta(X_i) = \frac{1}{1 + e^{-X_i^T \theta}}.$$

This then gives us a log likelihood function

$$l(\theta) = \sum_{i=1}^n T_i \log(\pi_\theta(X_i)) + (1 - T_i) \log(1 - \pi_\theta(X_i))$$

A common approach to estimating $\hat{\theta}$ is to do maximum likelihood, which requires solving:

$$\frac{\partial}{\partial \theta} l(\theta) = \frac{1}{n} \sum_{i=1}^n \left(\frac{T_i}{\pi_\theta(x_i)} - \frac{1 - T_i}{1 - \pi_\theta(X_i)} \right) \pi'_\theta(x_i) = 0 \tag{21}$$

Notice: this is exactly the IPW estimator multiplied by a $\pi'_\theta(x_i)$. The above is the sample version of the following moment condition:

$$E\left(\left(\frac{T_i}{\pi_\theta(x_i)} - \frac{1 - T_i}{1 - \pi_\theta(X_i)} \right) \pi'_\theta(x_i) \right) = 0$$

We can add another moment condition if we consider the values we are ultimately trying to estimate using the IPW estimator - namely $E(Y_i(1)) - E(Y_i(0)) = \mu_1 - \mu_0$. We would like the following to hold (giving the theoretical version and the sample replacement).

$$\mathbb{E}\left(\frac{T_i Y_i}{\pi_\theta(X_i)} \right) = \mu_1 \qquad \rightarrow \qquad \frac{1}{n} \sum_{i=1}^n \frac{T_i Y_i}{\pi_\theta(X_i)} = \mu_1$$

[46] See (van der Vaart, 1998, Ch.4) for a reference on this. Not required for this course.

$$\mathbb{E}\left(\frac{(1-T_i)Y_i}{1-\pi_\theta(X_i)}\right) = \mu_1 \quad \rightarrow \quad \frac{1}{n}\sum_{i=1}^{n}\frac{(1-T_i)Y_i}{1-\pi_\theta(X_i)} = \mu_0$$

This gives us three conditions and three unknowns (θ, μ_1, μ_0). We can solve this system to obtain estimators for all three parameters. The solution is really just solving the first condition and plugging it into the second and third to get μ_0, μ_1...exactly what we would be doing if we took an MLE approach to estimating $\hat{\theta}$ and plugged it into the IPW estimator in Section 8.2.2. See Slide 15 for the Hajek estimator version of this. The nice thing about formulating the problem this way is that if our estimator is the solution to a MoM set-up, we can apply standard MoM asymptotic variance results to describe the asymptotic normality and variance of $(\hat{\theta}, \hat{\mu}_1, \hat{\mu_0})$ and could, for example, apply the Delta Method to obtain a variance for $\hat{\mu}_1 - \hat{\mu}_0$.

Bootstrapping: in practice, sometimes one of the most accessible variance estimation approaches is the boot strap. The basic procedure here is to resample our data B times with replacement and for each sample (a) re-fit the propensity score model or other method of calculating the weights (b) re-calculate the estimator. This gives us B copies $\hat{\tau}^{(k)}$ for $k = 1, ..., B$ and we can use their variance and quantiles to estimate standard errors and calculate confidence intervals.

8.3 Extensions: Other Weighting Approaches

A general problem with weighting is that weights can end up being non-uniform and extreme weights can create instability and hence high variance. Sometimes, simply normalizing the propensity score weights is not enough to fix this issue. For example, if the propensity score model is overfit to the data, we may end up getting estimated probabilities of treatment of mostly either 0 or 1! There are a few different frameworks for thinking about **optimal weighting**. One key idea is that if covariate balance is ultimately the thing we are aiming for, we might pick our weights to target that directly. The *true* propensity scores do achieve balance (Section 8.2.1), but our estimates of them may not be optimized to fulfill this property.

8.3.1 Covariate Balancing Propensity Score (CBPS)

As explained in Section 8.2.4, one way of viewing the IPW estimator is as solving for θ in the following equation, which can be viewed as the empirical version of a moment condition

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{T_i}{\pi_\theta(x_i)}\right)\frac{\partial}{\partial\theta}\pi_\theta(x_i) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1-T_i}{1-\pi_\theta(X_i)}\right)\frac{\partial}{\partial\theta}\pi_\theta(x_i) \tag{22}$$

The CBPS approach solves for θ in our propensity score model $\pi_\theta(X_i)$ by solving the following theoretical moment conditions

$$\mathbb{E}\left(\frac{T_i}{\pi_\theta(X_i)}f(X_i)\right) = \mathbb{E}\left(\frac{1-T_i}{1-\pi_\theta(X_i)}f(X_i)\right) = \mathbb{E}(f(X_i))$$

where $f(X_i)$ can be a vector. When replaced with their sample analogues, these become

$$\frac{1}{n}\sum_{i=1}^{n}\frac{T_i}{\pi_\theta(X_i)}f(X_i) = \frac{1}{n}\sum_{i=1}^{n}\frac{1-T_i}{1-\pi_\theta(X_i)}f(X_i) = \frac{1}{n}\sum_{i=1}^{n}f(X_i)$$

and hence we have a (generalized) method of moments estimator. Here, $f(X_i)$ is a generalization of the propensity score case where $f(X_i) = \frac{\partial}{\partial\theta}\pi_\theta(X_i)$. The key point is that the standard maximum likelihood approach to estimating propensity scores is actually optimizing to balance the weighted mean first derivative of the propensity score model and it isn't clear this is useful or optimal. By comparison, $f(X_i) = X_i$ (a vector of p covariates) might result in conditions that reflect exactly the kinds of quantities we'd calculate to check balance. When the estimand is the ATE, the optimal choice of f is

$$f(X_i) = \pi_\theta(X_i)\mu_0(X_i) + (1-\pi_\theta(X_i))\mu_1(X_i)$$

this has double robustness properties (see next section) and minimizes the asymptotic variance when the propensity score is correct (Fan et al., 2023). This does involve the outcome model, which one would need to specify. The presence of the outcome model reflects the idea ultimately, we want to balance the covariates which have some relationship to the outcome. Note also that including an intercept term $f(X_i) = 1$ creates a criteria of the form

$$\frac{1}{n}\sum_{i=1}^{n}\frac{T_i}{\pi_\theta(X_i)} = \frac{1}{n}\sum_{i=1}^{n}\frac{1-T_i}{1-\pi_\theta(X_i)} = 1$$

which creates normalization of the treatment weights and the control weights. For example, for the treatment group, we have weight $w_i = \frac{1}{n\pi_\theta(X_i)}$ with $\sum_{i=1}^{n}T_i w_i = 1$. It is a good to include this because it leads to less extreme weights and smaller variance.

CBPS can be implemented using the `CBPS` function in the `CBPS` package in `R`, which has an argument for specifying whether the ATT or ATE is of interest.

8.3.2 Calibration Methods

CBPS still involves specifying a propensity score model. Calibration methods avoid modeling propensity scores directly and focus on balancing covariates, which some argue is really the goal of weighting. Some methods also incorporate constraints to avoid making weights too extreme, which inflates variance. These methods consist of some objective function and constraints together with an approach for optimizing that objective function.

Entropy Balancing (Hainmueller, 2012): for the ATT involves the following optimization: Let w^* be a length n_0 vector of weights. Find

$$w^* = \arg\min_w \sum_{i:T_i=0} w_i \log\left(\frac{w_i}{q_i}\right)$$

subject to

- positivity $w_i \geq 0$
- normalization $\sum_{i:T_i=0} w_i = 1$
- balancing condition $\sum_{i:T_i=0} w_i f(X_i) = \frac{1}{n_1} \sum_{i:T_i=1} f(X_i)$

for some choice of function f and some choice of target weight distribution q_i. Typically, $q_i = \frac{1}{n_0}$ (uniform distribution) and we minimizing the the KL divergence between the weights (which can be seen as forming some discrete distribution over indices $1, ..., n_0$) and the target distribution. We can easily flip the above specification to calculate weights for the ATC case.

This optimization problem is convex, making it more feasible to implement. The balancing condition ensures that the resulting solution will achieve balance on the $f(X_i)$ function of interest (this can just be $f(X_i) = X_i$, where X_i is a vector of p covariates). A disadvantage is that this method does not necessarily solve the extreme weights problem. There is nothing in the optimization problem that prevents them from being extreme. This method can be implemented in `R` using the `ebalance` function in the `ebal` package. By default, it calculates the weights for the ATT case.

Stable Weights (Zubizarreta, 2015) explicitly targets balancing but also tries to avoid weights being too extreme. The optimization problem for the ATT case is to minimize the variance of the weights:

$$w^* = \arg\min_w \|w - \bar{w}\|_2^2 = \arg\min_w \sum_{i=1}^{n} (w_i - \bar{w})^2$$

subject to

- positivity $w_i \geq 0$
- normalization $\sum_{i:T_i=0} w_i = 1$
- balancing condition $\left|\sum_{i:T_i=0} w_i X_{ij} = \frac{1}{n_1} \sum_{i:T_i=1} X_{ij}\right| \leq \delta_j$ for covariates $j = 1, ..., p$.

δ_j represents a pre-specified level of covariate balance for covariate j. We do not require balance to hold exactly as in entropy balancing. Again this call all be easily flipped for the ATC case. The resulting optimization problem is a quadratic convex programming problem. Stable weights is implemented in the `sbw` package in `R`.

8.3.3 Overall

So which weighting method should you use? There is no overall consensus. Propensity scores are, in theory, the right thing to do – if you truly knew $\pi(X_i)$, then the IPW estimator is unbiased. Propensity scores are also appealing because they are interpretable and you might be able to reason about the plausibility that $\pi(X_i)$ actually describes the probability of treatment. By comparison, weights obtained via calibration methods or even via CBPS can be harder to interpret. These weights do, however, directly target balance, while propensity scores are not directly estimated with this objective in mind. Calibration methods can also help with making weights less extreme, which is an issue for variance. Note, however, that while in IPW, there are no assumptions about the outcome model of Y_i given X_i, CPBS and the calibration methods above implicitly do involve the outcome model via the choice of $f(X_i)$ and the quality of the results can depend on the outcome model. For example, if Y_i has a linear relationship to X_i, balancing on $f(X_i) = X_i$ can work well, but if Y_i is a function of X_i^2, it would be better to balance X_i^2 values.

8.3.4 Checking Covariate Balance for Weighting

Even for methods which optimize to achieve covariate balance explicitly, it is important to always check covariate balance once some weighting method has been applied. The question is whether, with weights applied, the treatment and control group look similar. When estimating the ATT, the weighting version of the smd formula introduced in the section on matching (Equation (20)) for checking balance is:

$$\text{smd}_{j,treat} = \frac{\bar{X}_{j1} - \bar{X}_{j0}}{S} \tag{23}$$

where

- $\bar{X}_{j1}$ is the *unweighted* mean for the treatment group
- $\bar{X}_{j0} = \frac{\sum_{i:T_i=0} w_i X_{ij}}{\sum_{i':T_{i'}=0} w_{i'}}$ is the weighted mean of covariate j in the control group
- $S = \sqrt{\frac{1}{n_1-1} \sum_{i=1}^{n} T_i (X_{ij} - \bar{X}_{j1})^2}$ is the standard deviation of covariate j within the treatment group (not using weights)

A similar version can be calculated when our focus is on estimating the ATC – there we would calculate a weighted mean for the treatment group and an unweighted mean for the control group. For the ATE, we'd calculate weighted means for both groups and standardize by a pooled variance or sum of variances across treatment and control groups. The key idea is that we want the weights to make the control group look like the treatment group or vice versa in terms of pre-treatment covariates. We standardize by the level of noise present in whichever group we are interested in estimating causal effects for.

For example, if there are few older people in the treatment group and many in the control group, we might hope that with weighting to upweight older people in the treated group (e.g., using their small propensity score) and downweight them in the control group, the weighted mean age for treatment and control will be close. Again, it makes sense introduce some standardization here so that we can evaluate resulting differences on a common scale.

8.4 Doubly Robust Estimation

Module 7 Section 7.2 describes a regression estimator which requires us to correctly specify a regression of Y on T and X.

$$\hat{\tau}_{reg} = \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)$$

Module 8 Section 8.2.2 describes IPW weighting estimators which requires a correctly specified propensity model for $\pi(X_i)$.

$$\hat{\tau}_{IPW} = \frac{1}{n}\sum_{i=1}^{n} \frac{T_i Y_i}{\hat{\pi}(X_i)} - \frac{(1-T_i)Y_i}{1-\hat{\pi}(X_i)}$$

If the regression or propensity score model is misspecified, this can introduce bias into the respective estimators! This raises a few questions:

1. **Q**: In the well-specified case, which is better? **A**: the outcome regression has lower variance
2. **Q**: Which is easier to specify correctly? **A**: propensity score modeling can be easier because we may have more knowledge of how treatment was assigned...but specifying a correct model can be difficult either way.
3. **Q**: Do we really have to choose between them? **A**: NO! Because we have doubly robust estimators!

The classic **doubly Robust** estimator for this setting is the **Augmented Inverse Probability Weighting (AIPW) estimator**, which incorporates both the propensity score model and regression model in such a way that as long as *one* of the two models is correctly specified, the estimator is consistent for the estimand. This somewhat magical property gives us *two* chances to be right! There is some variance cost in that this estimator still has higher variance than a correctly specified outcome model, but the double robustness is an appealing property. It is also still an efficient estimator in some sense in that "it has the smallest asymptotic variance among estimators that are consistent when the propensity score model is correct" (Slide 16).

8.4.1 AIPW Estimator of the ATE

There are two important, algabraically equivalent ways of writing this estimator.

$$\hat{\tau}_{DR,ATE} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{T_i Y_i}{\hat{\pi}(X_i)} - \frac{T_i - \hat{\pi}(X_i)}{\hat{\pi}(X_i)}\hat{\mu}_1(X_i)\right] - \left[\frac{(1-T_i)Y_i}{1-\hat{\pi}(X_i)} - \frac{T_i - \hat{\pi}(X_i)}{1-\hat{\pi}(X_i)}\hat{\mu}_0(X_i)\right]$$

$$\hat{\tau}_{DR,ATE} = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{\mu}_1(X_i) + \frac{T_i(Y_i - \hat{\mu}_1(X_i))}{\hat{\pi}(X_i)}\right] - \left[\hat{\mu}_0(X_i) + \frac{(1-T_i)(Y_i - \hat{\mu}_0(X_i)}{1-\hat{\pi}(X_i)}\right]$$

The forms above give some intuition for why this estimator is doubly robust. In the first line above, if the propensity scores model is well-specified, then the $T_i - \hat{\pi}(X_i)$ in the $\frac{1}{n}\sum_{i=1}^{n}\frac{T_i - \hat{\pi}(X_i)}{\hat{\pi}(X_i)}\hat{\mu}_1(X_i)$ and $\frac{1}{n}\sum_{=1}^{n}\frac{T_i - \hat{\pi}(X_i)}{1-\hat{\pi}(X_i)}\hat{\mu}_0(X_i)$ terms will tend to cancel so that those terms are ≈ 0 and we are left exactly with the IPW estimator, which is then a consistent estimator for the true ATE (assuming unconfoundedness, overlap). In the second line above, if the regression model is correct, then the $\frac{1}{n}\sum_{i=1}^{n}\frac{T_i(Y_i - \hat{\mu}_1(X_i))}{\hat{\pi}(X_i)}$ terms and its parallel term in the second part will be ≈ 0 and we get exactly the regression estimator, which is then a consistent estimator. Of course, if neither is well-specified, this estimator can still be biased.

More formally, I sketch the argument for double robustness below using the first formulation of $\hat{\tau}_{DR}$. Note we assume throughout that unconfoundedness and overlap hold, that we have i.i.d. data, and that any relevant expectations are finite.

1. By the Law of Large Numbers, $\hat{\tau}_{DR}$ converges to the expected value of the quantity within the sum
2. Suppose that $\pi_\theta(X_i)$ and $\mu_{t,\beta}(X_i)$ are the chosen propensity score and outcome models and that estimators $\hat{\theta}$ and $\hat{\beta}$ are consistent for their true values: $\hat{\theta} \xrightarrow{p} \theta$, $\hat{\beta} \xrightarrow{p} \beta$.[47] Then by applying some convergence arguments, we can argue that

$$\hat{\tau}_{DR,ATE} \xrightarrow{p} \mathbb{E}\left[\left[\frac{T_i Y_i}{\pi_\theta(X_i)} - \frac{T_i - \pi_\theta(X_i)}{\pi_\theta(X_i)}\mu_{1,\beta}(X_i)\right] - \left[\frac{(1-T_i)Y_i}{1-\pi_\theta(X_i)} - \frac{T_i - \pi_\theta(X_i)}{1-\pi_\theta(X_i)}\mu_{0,\beta}(X_i)\right]\right] \tag{24}$$

   Where note here that we may have $\mathbb{E}(T_i|X_i) \neq \pi_\theta(X_i)$ and $\mathbb{E}(Y_i|X_i, T_i = t) \neq \mu_{t,\beta}(X_i)$ if models are mis-specified.

[47] Note: estimators can be consistent even if a model is mis-specified. For example, in linear regression, if the true model is $Y_i = X_i^2\delta + \eta_i$ and we fit $Y_i = X_i\beta + \epsilon_i$, we still have a true unknown value of β (a true best linear predictor), even if it is not δ.

3. Consider just the first term. We will show that as long as one model is correctly specified, this is equal to $\mathbb{E}(Y_i(1))$.

$$\begin{aligned}
\mathbb{E}\left[\frac{T_iY_i}{\pi_\theta(X_i)} - \frac{T_i - \pi_\theta(X_i)}{\pi_\theta(X_i)}\mu_{1,\beta}(X_i)\right] &= \mathbb{E}\left[\frac{T_iY_i(1)}{\pi_\theta(X_i)} - \frac{T_i - \pi_\theta(X_i)}{\pi_\theta(X_i)}\mu_{1,\beta}(X_i) \pm \frac{T_i - \pi_\theta(X_i)}{\pi_\theta(X_i)}Y_i(1)\right] \\
&= \mathbb{E}\left[Y_i(1)\left(\frac{T_i}{\pi_\theta(X_i)} - \frac{T_i - \pi_\theta(X_i)}{\pi_\theta(X_i)}\right) + \frac{T_i - \pi_\theta(X_i)}{\pi_\theta(X_i)}(Y_i(1) - \mu_{1,\beta}(X_i))\right] \\
&= \mathbb{E}\left[Y_i(1)\right] + \mathbb{E}\left[\frac{T_i - \pi_\theta(X_i)}{\pi_\theta(X_i)}(Y_i(1) - \mu_{1,\beta}(X_i))\right]
\end{aligned}$$

Above, the **strategy** is to first try to isolate $Y_i(1)$ and then deal with whatever term remains. Our goal is now to show that this term is 0 as long as one model is well-specified.

4. **Scenario 1:** Suppose the propensity score model is correctly specified. Then using the law of total expectation and unconfoundedness,

$$\begin{aligned}
\mathbb{E}\left[\frac{T_i - \pi_\theta(X_i)}{\pi_\theta(X_i)}(Y_i(1) - \mu_{1,\beta}(X_i))\right] &= \mathbb{E}\left[\mathbb{E}\left[\frac{T_i - \pi_\theta(X_i)}{\pi_\theta(X_i)}(Y_i(1) - \mu_{1,\beta}(X_i))|X_i\right]\right] \\
&= \mathbb{E}\left[\left(\frac{\mathbb{E}[T_i|X_i]}{\pi_\theta(X_i)} - 1\right)(\mathbb{E}(Y_i(1)|X_i) - \mu_{1,\beta}(X_i))\right] \\
&= \mathbb{E}\left[(1-1)(\mathbb{E}(Y_i(1)|X_i) - \mu_{1,\beta}(X_i))\right] \\
&= 0
\end{aligned}$$

5. **Scenario 2:** Suppose the outcome model is correctly specified. Then we can apply the same calculations as above but now, we also use unconfoundedness to argue $\mathbb{E}(Y_i(1)|X_i) = \mathbb{E}(Y_i(1)|X_i, T_i = 1)$

$$\begin{aligned}
\mathbb{E}\left[\frac{T_i - \pi_\theta(X_i)}{\pi_\theta(X_i)}(Y_i(1) - \mu_{1,\beta}(X_i))\right] &= \ldots = \mathbb{E}\left[\left(\frac{\mathbb{E}[T_i|X_i]}{\pi_\theta(X_i)} - 1\right)(\mathbb{E}(Y_i(1)|X_i) - \mu_{1,\beta}(X_i))\right] \\
&= \mathbb{E}\left[\left(\frac{\mathbb{E}[T_i|X_i]}{\pi_\theta(X_i)} - 1\right)(\mathbb{E}(Y_i(1)|X_i, T_i = 1) - \mu_{1,\beta}(X_i))\right] \\
&= \mathbb{E}\left[\left(\frac{\mathbb{E}[T_i|X_i]}{\pi_\theta(X_i)} - 1\right) * 0\right] \\
&= 0
\end{aligned}$$

6. The argument for showing the second term in Equation 24 has expected value $E(Y_i(0))$ is similar.

Doubly-robust estimators also exist for the ATC and ATT. Stat286 students will consider the ATC doubly-robust estimator in the Module 8 problem set.

8.5 Further Resources

1. Online book with a short chapter on matching that includes links to further resources and some notes on relevant R packages. https://bookdown.org/paul/applied-causal-analysis/matching.html
2. Imbens and Rubin (2015) Chapters 12-15, 18-19

9 Module 9: Causal Mechanisms

Big Picture: if we learn that a treatment has a causal effect on an outcome, the next question might be, "well, how?" Researchers and policymakers often care about the *mechanisms* through which a treatment has an effect on an outcome. For example, knowing these mechanisms might inform whether a treatment is likely to be generalizable to other contexts and populations. In this module, we formalize the notion of a mediators in terms of potential outcomes and examine how we might identify their effects for experimental or observational studies.

Running Example: Health Insurance and Surgery. Imagine we are interested in whether people who get health insurance (T) have improved health outcome (Y) and want to know whether that is because they are more likely to get a certain surgery ($M = m$).

9.1 Defining Causal Mediators and Related Estimands

As usual, we will focus on binary treatment $T \in \{0, 1\}$ and are interested in the causal effect of T on an outcome Y. Now, however, we introduce mediator M which takes values $m \in \mathcal{M}$. We then think of both potential mediator values $M_i(t)$ and potential outcomes as a function of both treatment and mediator $Y_i(t, m)$. As always in this course, we assume **consistency**, so $M_i = M_i(T_i)$ and $Y_i = Y_i(T_i, M_i)$. The DAG representation of this set-up is

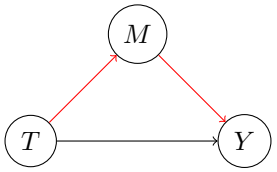


Figure 26: Simple mediation DAG

The path $T \to M \to Y$ represents Y depending **indirectly** on treatment via a mediator that is affected by treatment. The path $T \to Y$ represents some remaining **direct** dependence on Y. This path reflects all other, possibly unobserved and unknown mediators. That is, in principle, if we added enough mediators, we should be able to remove the direct $T \to Y$ arrow. In thinking about the different effects at play here, an important distinction is between **controlled effects** and **natural effects**. Controlled effects reflect what would happen if we could intervene to fix the mediator and treatment value (in DAG terminology, do(t) and do(m) operations). They involve comparisons of $Y_i(t, m)$ for $t = 0, 1$ at different $M = m$ values. Natural effects reflect what would happen if we intervene on treatment (do(T)) and let M take whatever value it *naturally* takes under different treatment conditions. We saw this, for example, in Module 5 where we considered the effect that an encouragement had on treatment and treatment subsequently had on outcome and considered quantities like $Y_i(Z_i, T_i(Z_i))$.

Terminology note: moderation vs mediation: generally, when we talk about a variable *moderating* a causal effect we mean pre-treatment covariates and are concerned with how treatment effects may vary as a function of these covariates. *Mediation* instead refers to the way treatment affects the overall outcome of interest via its effect on post-treatment variables that have an effect on outcome.

Perspective note: in this section, we approach causal mediation from the perspective of a treatment *causing* an intermediary variable, which affects the outcome. Another approach, which we do not focus on, is to think about decomposing the treatment into different components. For example, a health intervention treatment could encompass different things and it could be that some aspects of the treatment (e.g., being encouraged to walk more) matter while others (e.g., taking vitamin D) do not matter for a given health outcome of interest.

9.1.1 Controlled Direct Effects (CDE)

The individual and average controlled direct effects are defined as for each $m \in \mathcal{M}$

$$\xi_i(m) = Y_i(1, m) - Y_i(0, m)$$

$$\bar{\xi}(m) = \mathbb{E}(Y_i(1, m) - Y_i(0, m))$$

This is the direct effect on the outcome when we hold the mediator constant at some value m (essentially intervening to fix it, $do(m)$ in DAG notation). The idea is that a remaining difference in Y_i must then come from something other than the mediator. The value of $\xi_i(m)$ can vary with the value of m (an interaction effect).

If m fully accounts for the causal effect of T on Y, then we could have $\xi_i(m) = 0$ for all m (or $\bar{\xi}(m) = 0$ if this is on average) but this does not mean that treatment has no effect on outcome! If we estimate the CDEs (see below) and they are all ≈ 0 while our estimate of the total treatment effect (ignoring mediator) is non-zero, this suggests the mediator explains a lot of the effect of the treatment. Note, however, that this quantity is not directly capturing anything about how much the mediator affects Y or about which values mediators tend to take in practice (see natural effects).

Running Example: This corresponds to fixing surgery status (e.g., to no surgery) and looking at the difference in heart disease potential outcome with or without insurance. This effect is not "natural" because it may be that individual i who got health insurance would actually have gotten the surgery ($Y_i = Y_i(t = 1, m = 1)$) while we are asking about a world in which the person takes a different action. One can imagine a scenario in which few people with $T_i = 1$ have $M_i = 0$ and few with $T_i = 0$ have $M_i = 1$ so that effects involving $Y_i(1, 0)$ and $Y_i(0, 1)$ do not reflect real-world dynamics.

Are there "Controlled Indirect Effects?" You might wonder if we could also define a controlled indirect effect. We *could* write down the quantity $Y_i(t, m) - Y_i(t, m')$ but we do not call this an indirect effect because it just represents a direct causal effect of the mediator if we were to fix m to different values – possibly an effect *moderated* by the level of treatment. That is, once we fix $T = t$ and do not look at how t induces certain $M = m$ values, we have essentially made M part of our treatment. We could identify $E(Y_i(t, m) - Y_i(t, m'))$, e.g. by randomizing T and M and treating (t, m) pairs as our treatments levels, but that no longer involves M as a mediator. It is different from the situation where treatment has an effect on M_i. That is the situation we consider with natural effects.

9.1.2 Natural Direct Effects (NDE)

The individual and average natural direct effects are

$$\zeta_i(t) = Y_i(1, M_i(t)) - Y_i(0, M_i(t))$$

$$\bar{\zeta}(t) = \mathbb{E}(Y_i(1, M_i(t)) - Y_i(0, M_i(t)))$$

Here, we are again fixing the value of the mediator but to whatever value it *naturally* takes as a function of treatment t. For example, if $t = 1$, we have

$$\zeta_i(1) = Y_i(1, M_i(1)) - Y_i(0, M_i(1))$$

Hence we are looking at the outcome under treatment and whatever value the mediator would take under treatment compared to the value under control and *whatever value the outcome would have taken under treatment.*

Running Example: imagine a treated patient who gets the surgery. We can then (hypothetically) compare her actual outcome under treatment and surgery to what would have happened if she had not been treated but still got the surgery. This tells us about the direct effect of treatment on her outcome. If this patient is an always-taker of surgery $M_i(1) = M_i(0) = 1$, then her natural direct effect is identical for $t = 0, 1$. If, without health insurance, she would not have gotten surgery, the natural direct for $t = 0$ compares outcomes while fixing $M_i(t) =$ no-surgery and that effect could be different from the $t = 1$ case.

9.1.3 Natural Indirect Effect (NIE)

Note: also known as causal mediation effect.

The individual and average natural indirect effects are

$$\delta_i(t) = Y_i(t, M_i(1)) - Y_i(t, M_i(0))$$

$$\bar{\delta}(t) = \mathbb{E}(Y_i(t, M_i(1)) - Y_i(t, M_i(0)))$$

Here, we fix the treatment status and compare the difference in outcome from changing the mediator to the *possibly*-different values it *naturally* takes under treatment or control. If $M_i(1) = M_i(0)$ (treatment has no effect on mediator), then automatically the NIE is 0. Otherwise, it may or may not be. Another way of putting this is that it is the causal effect of M on Y

that would be induced by T shifting from 0 to 1 but holding T fixed at a certain level. Note that this is **not** the same as the causal effect of the mediator. In terms of randomized experiments, we are not intervening on M to randomize it – we only randomize T.

Running Example: we can imagine treating patient i so that $T_i = 1$ (gets health insurance). Suppose that under treatment, this patient gets the surgery. If this patient would have gotten the surgery regardless of treatment, then there is no indirect effect of treatment on outcome via an effect on surgery. But if this patient would not have gotten surgery if not given health insurance, then it is meaningful to ask whether the surgery then drove a difference in final outcome. We therefore consider whether there is a difference in outcome under treatment + surgery vs under treatment + no surgery. Intuitively, if these outcomes are different, then treatment is at least partially having an effect on outcome via its effect on surgery. We fix treatment level because it may be that, due to a direct effect (other mechanisms), outcomes with health insurance are systematically different from outcomes without health insurance. Again there could be interaction dynamics where, for example, given you have health insurance (and get surgery), not getting surgery would have hurt your outcome but given you do not have health insurance (and do not get surgery), getting surgery would have hurt your outcome (or vice versa).

9.1.4 Total Effect

Notice that each of the effects described above includes $Y_i(t, m)$ that are not just unobserved because treatment is assigned to $t' \neq t$ but unobserved because that mediator value might never actually happen for that treatment. For example, we might write $Y_i(0, m)$ for an individual but have $M_i(0) \neq m$ so that even if $T_i = 0$, we would not observe that potential outcome. The **Total Effect** reflects what actually happens when $T_i = 0$ vs $T_i = 1$:

$$\tau = Y_i(1, M_i(1)) - Y_i(0, M_i(0))$$

$$\bar{\tau} = \mathbb{E}(Y_i(1, M_i(1)) - Y_i(0, M_i(0)))$$

9.1.5 Decomposition of Total Effect

The natural effect perspective has the advantage of allowing a decomposition of the treatment effect into direct and indirect effects (we cannot do this for controlled direct effect). There are a few different ways to write this:

$$\tau = \delta_i(t) + \zeta_i(1-t) = \frac{1}{2} \sum_{t=0,1} \delta_i(t) + \zeta_i(t)$$

The right form tells us that the total effect is the average direct effect + the average indirect effect over $t = 0, 1$.

Proof: *Decomposition of Total Effect.*

$$\begin{aligned}
\delta_i(1) + \delta_i(0) + \zeta_i(1) + \zeta_i(0) &= Y_i(1, M_i(1)) - Y_i(1, M_i(0)) \\
&\quad + Y_i(0, M_i(1)) - Y_i(0, M_i(0)) \\
&\quad + Y_i(1, M_i(1)) - Y_i(0, M_i(1)) \\
&\quad + Y_i(1, M_i(0)) - Y_i(0, M_i(0)) \\
&= 2(Y_i(1, M_i(1)) - Y_i(0, M_i(0)) \\
&= 2\tau
\end{aligned}$$

Note also that

$$\delta_i(t) + \zeta_i(1-t) = Y_i(t, M_i(1)) - Y_i(t, M_i(0)) + Y_i(1, M_i(1-t)) - Y_i(0, M_i(1-t))$$

If we plug in $t = 1$, we get

$$Y_i(1, M_i(1)) - Y_i(1, M_i(0)) + Y_i(1, M_i(0)) - Y_i(0, M_i(0)) = \tau$$

and similarly for $t = 0$

□

9.2 Identifying and Estimating the CDE

As usual, identification of the quantities described above depends on some assumptions. In some cases, these are versions of the standard assumptions we have grown used to in previous Modules, but there are also some new ones. To consider how conditioning on covariates can help and what can go wrong with post-treatment ones, let X be some **pre-treatment** covariates X and let Z be additional **post-treatment** variable(s) which could themselves be other mediators caused by T and having an effect on Y.

Running Example: extend the health insurance example to include age group X and an indicator for receiving some form of preventative care Z that may or may not happen after health insurance (T).

We will consider how to identify the CDE even in the presence of an additional post-treatment mediator Z that can affect M and Y as in the following DAG:

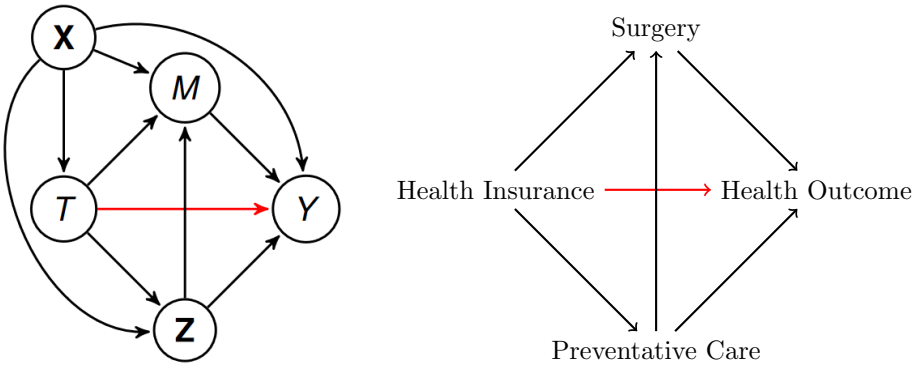


Figure 27: **Left:** DAG from Slide 7. Note that none of the arrows coming from X are essential for any of the arguments below. X is a pre-treatment confounder while Z is a post-treatment confounder. **Right**: hypothetical running example, without any X for simplicity

Note that it is fine for there to be no Z – in that case, it is simply dropped from the conditioning in assumption 2 below and the identification argument simplifies.[48] However, we consider the more general case to show how to identify the CDE if Z is present.

Assumptions: To identify the CDE, we require **sequential unconfoundedness**, which consists of the following:

1. **Treatment unconfoundedness:** $\{Y_i(t,m), M_i(t')\} \perp\!\!\!\perp T_i | X_i = x \quad \forall \quad t, t', x$
2. **Mediator unconfoundedness:** $Y_i(t,m) \perp M_i(t) | X_i = x, T_i = t, Z_i = z \quad \forall \quad t, x, z$

The first assumption is our standard unconfoundedness assumption: conditional on the pre-treatment covariates, treatment is assigned at random, independent of all the potential outcomes involved. The second assumption says that if we now condition on Z and T too (everything that is causally prior to M), then there are no other, unaccounted for factors that affect both M and the outcome. Note here that we are assuming here that Z is something we measure. If Z is unmeasured, then it plays the role of an unmeasured confounder and leads to a violation of Assumption 2! Note that it would be acceptable for there to be an unmeasured variable U in the DAG above that points only to M – some other factor affecting which mediator people get without any direct link to the outcome.

Running Example: The first assumption says that within age group (X), health insurance is assigned at random regardless of how people would be benefited or hurt by or it or whether people would or would not get the surgery. The latter assumption says that within age group ($X = x$), within the group that got (or did not get) health insurance ($T = 1$), for the patients who got (or did not get) preventative care ($Z = 1$), there is no way in which surgery (M) is systematically going to those who would most (or least) benefit from it. This would be violated if, for example, $X = x, T = 1, Z = 1$ people who attend hospital 1 are both more likely to get the surgery and have baseline lower health as compared to people within that group who attend hospital 2. Note that even if Assumptions 1 and 2 hold, it could still be that surgery leads to better outcomes ($Y_i(t, m = 1) > Y_i(t, m = 0)$) so that observationally, we do see a correlation between M_i and Y_i given $X_i = x, T_i = t, Z_i = z$. But as we've discussed since the start of the course, it is key that people are not being assigned surgery in some way related to their potential outcome strata.

[48] If $Z \to M$ were not present, then we could collapse it into the $T \to Y$ arrow and ignore it in the arguments below. If $Z \to Y$ were not present, we could similarly ignore Z as a variable along the $T \to M$ pathway. We are only concerned here with there being a Z that affects both.

9.2.1 Identification

The above assumptions allow the following identification argument for $\mathbb{E}(Y_i(t,m))$:[49]

$$\begin{aligned}
\mathbb{E}(Y_i(t,m)) &= \mathbb{E}(\ \mathbb{E}(Y_i(t,m)|X_i)) && \text{(Law of Tot. Exp.)}\\
&= \mathbb{E}(\ \mathbb{E}(Y_i(t,m)|T_i=t,X_i)) && \text{(Assumption 1)}\\
&= \mathbb{E}(\ \mathbb{E}(\ \mathbb{E}(Y_i(t,m)|T_i=t,X_i,Z_i)|T_i=t,X_i)) && \text{(Law of Tot. Exp)}\\
&= \mathbb{E}(\ \mathbb{E}(\ \mathbb{E}(Y_i(t,m)|T_i=t,X_i,Z_i,M_i=m)|T_i=t,X_i)) && \text{(Assumption 2)}\\
&= \mathbb{E}(\ \mathbb{E}(\ \mathbb{E}(Y_i|T_i=t,X_i,Z_i,M_i=m)|T_i=t,X_i)) && \text{(Consistency)}\\
&= \sum_{x,z}\mathbb{E}(Y_i|T_i=t,X_i=x,Z_i=z,M_i=m)\Pr(Z_i=z|T_i=t,X_i=x)\Pr(X_i=x) && \text{(By def, if } X,Z \text{ discrete)}
\end{aligned}$$

The argument for $E(Y_i(0,m))$ is identical, except with $T_i=0$. Note: we need this more complex argument because Z is a post-treatment variable. Note that it would not work to simply regress Y_i on X,Z for $T_i=1$ and $M_i=m$ and average these as we did in Module 7 Section 7.2 because Z is a post-treatment variable. Doing so would create **post-treatment bias**. Instead, the identification equality above indicates we need to model the distribution of Z given T and X in addition to modeling the outcome.

9.2.2 Estimation

Given the above identification equality, we require models of the following:

- $E(Y_i|T_i=t,X_i,Z_i,M_i=m)$ an outcome regression of Y_i on X_i and Z_i within the $M_i=m,T_i=1$ group
- $\Pr(Z_i=z|T_i=t,X_i=x)$ a regression of Z_i on X_i within the $T_i=1$ group
- $\Pr(X_i=x)$ estimated via sample proportions.

Note that fitting these models can be difficult if X and/or Z are high-dimensional.

There are a few approaches to then using these models.

1. **Directly use identification formula with plug-in**: if your models have a closed form so that it is easy to calculate $\hat{\Pr}(Z_i=z|T_i=t,X_i=x)$ and for each z,t,x and similarly for the outcome model, then we could replace each term in the identification equality by the estimated models. If we model $P(X)$ by the empirical distribution, this is[50]

   $$\frac{1}{n}\sum_{i=1}^{n}\sum_{z}\hat{\mathbb{E}}(Y_i|T_i=1,X_i,Z_i=z,M_i=m)\widehat{\Pr}(Z_i=z|T_i=1,X_i)-\hat{\mathbb{E}}(Y_i|T_i=0,X_i,Z_i=z,M_i=m)\widehat{\Pr}(Z_i=z|T_i=0,X_i)$$

   Caution: while intuitive, this approach is not necessarily the optimal thing to do for reasons that relate to semi-parametric theory (not covered in this course)

2. **Directly use identification formula with prediction**: especially if we want to model $Z_i|T_i=t,X_i=x$ via a complicated model (e.g., Neural Network) that we cannot easily write down, we can instead *predict* Z_i given X_i and T_i by $\widehat{Z_i(1)}$ for $T_i=1$ and $\widehat{Z_i(0)}$ for $T_i=0$ and calculate

   $$\frac{1}{n}\sum_{i=1}^{n}\hat{\mathbb{E}}(Y_i|T_i=1,X_i,\widehat{Z_i(1)},M_i=m)-\hat{\mathbb{E}}(Y_i|T_i=1,X_i,\widehat{Z_i(0)},M_i=m)$$

   (again here we are using the empirical distribution of X)

[49] **Tip:** These calculations are a bit tedious, but notice we are repeating strategies we have seen throughout the course: To achieve identification, we want to go from potential outcomes to observational quantities. To do so, we want to apply consistency $Y_i=Y_i(T_i,M_i)$, but initially, we have $Y_i(t,m)$ and are not conditioning on $T_i=t$ or $M_i=m$. The strategy is to use some combination of the law of total expectation and unconfoundedness to bring these into the conditioning set. In each case we first take the things conditioned on as part of the unconfoundedness assumptions and use law of total expectation to bring them in. Then we can bring in the treatment or mediator using the assumption.

[50] The empirical distribution is just $P(X=x)=\frac{1}{n}\sum_{i=1}^{n}I(X_i=x)$. If we have i.i.d data, it is consistent for the true distribution, though for high dimensional X, it might require a high n for this approximation to be close.

9.2.3 Estimation Strategy Under Nested Mean Models with Additional Assumption

A third option for estimation is to make an additional **no-interaction** assumption:

$$\mathbb{E}(Y_i(t,m) - Y_i(t,m')|T_i = t, X_i, Z_i) = \mathbb{E}(Y_i(t,m) - Y_i(t,m')|T_i = t, X_i)$$

$Y_i(t,m) - Y_i(t,m')$ is the effect of the mediator on the outcome. In general, this effect could vary with the value of $Z_i = z$. That is, if Z has an effect on the outcome, then we could could write $Y_i(t,m,z)$ and have $Y_i(t,m,z) - Y_i(t,m',z) \neq Y_i(t,m,z') - Y_i(t,m',z')$. The above assumption says that, at least on average and conditional on $T_i = t$ and X_i, this does not happen. The average effect of the mediator on outcome is the same regardless of the value of Z_i - if there is heterogeneity in this effect, it is accounted for by conditioning on $T_i = t$ and $X_i = x$. What does this get us? It allows us to use an approach called **Structural Nested Mean Models (NNM)** (Slide 10), which we will explain here for the case of linear models.

The overall intuition for NNM is that, assuming there's no interaction, we are going to model the effect of the mediator on Y for each value of m and then use that model to adjust our observed Y_i. Via this adjustment, we hope to approximate what Y_i we would have observed holding m constant at some reference level or value m_0. Formally, the steps are:

1. **Regress** Y_i on observed M_i, T_i, X_i, Z_i without an interaction between M_i and Z_i

   $$\mathbb{E}(Y_i|M_i, T_i, X_i, Z_i) = \alpha_0 + \alpha_1 T_i + \alpha_2 M_i + \alpha_3^T X_i + \alpha_4^T Z_i$$

   as usual in regression, if M_i is discrete, we pick some reference level and otherwise m_0 is some reference value. α_2 represents the change in expected outcome for going from m_0 to the other level m.

2. **Calculate blip function**: using the model above and the unconfoundedness and no-interaction assumptions, we can calculate the "blip" function

   $$\begin{aligned}\gamma(t,m,X_i) &= \mathbb{E}[Y_i(t,m) - Y_i(t,m_0)|X_i] \\ &= \mathbb{E}[Y_i(t,m) - Y_i(t,m_0)|X_i, T_i, Z_i] \\ &= \alpha_2(m - m_0)\end{aligned}$$

   where if M is discrete, this would just be α_2. This function represents the average effect of the mediator on the outcome relative to the value at m_0.

3. **Adjust outcome** using blip function and then **regress** this on T and X.

   $$\mathbb{E}(Y_i - \gamma(T_i, M_i, X_i)|T_i, X_i) = \beta_0 + \beta_1 T_i + \beta_2^T X_i$$

   By subtracting $\alpha_2(m - m_0)$ (or, in practice, $\hat{\alpha}_2$), we adjust Y_i to its (estimated) expected value at $M_i = m_0$.

4. Take $\hat{\beta}_1$ as an estimate of the CDE with m held at m_0, i.e. of $\bar{\xi}(m_0)$.

A few comments: Z_i does still play a role above in the first regression step. It is still possible that the outcome is higher or lower for certain values of Z_i and we want to account for that. However, Z_i drops out in steps 2-3 because once we 'set' M_i to m_0 for everything, we essentially break the $Z \to M$ arrow in Figure 27.[51] The intuition for the later steps is that if, once we hold M_i constant at m_0, there is no further effect of treatment T_i, then the controlled direct effect is 0. This method works by trying to approximate that "if" condition and then seeing if there's a leftover effect.

See (Robins, 1994; Acharya et al., 2016) for more.

[51] As discussed in Section 7.6.5.

9.3 Identifying and Estimating NDE and NIE

Unlike for the CDE, to identify natural effects, we will need to assume there are no post-treatment confounders Z as appeared in in Figure 27. Instead, we assume the DAG in Figure 28, where we may have some pre-treatment covariate affecting all the other variables, and we may have other mediators buried in the $T \to Y$ arrow, but there are no other mediators that also affect M.

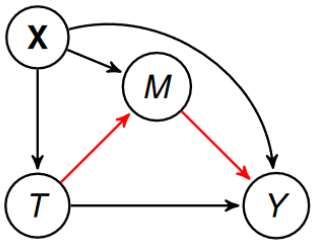


Figure 28: DAG from Slide 8

We now require one familiar unconfoundedness assumption and one new, tricky one:

Assumptions:

1. **Treatment unconfoundedness:** $\{Y_i(t,m), M_i(t')\} \perp\!\!\!\perp T_i | X_i = x \quad \forall \quad t, t', x$
2. **Cross-world counterfactual**: $Y_i(t', m) \perp M_i(t) | T_i = t, X_i = x$ for all t, x (including cases where $t' = t$ and $t' \neq t$)

The first assumption is the same as in the CDE case. The second assumption can be broken into two cases:

1. **Case of $t = t'$**: In this case, the assumption is identical to the CDE Assumption 2 in Section 9.2 only without the $Z_i = z$. We are saying there are no unobserved confounders that affect both M and Y.
2. **Case of $t \neq t'$**: This is the 'cross-world' part. It says that conditional on treatment status and $X_i = x$, the mediator variable in the conditioned on $T_i = t$ world is assigned independently of a person's potential outcomes for a different treatment status – i.e., in a counterfactual world different from the one we are conditioning on! In practice, as discussed in the example below, this essentially means there is no post-treatment confounder Z as in Figure 27. The reason this comes up in the natural effects world is that we are working with quantities like $Y_i(1, M_i(0))$ – the potential outcome under treatment $T_i = 1$ (world 1) and the mediator value the unit would have had under control (world 2). Unlike the quantity $Y_i(t, m)$ from the controlled effects world, we are now 'crossing worlds' within a single potential outcome!

   This assumption is sometimes controversial because it is more unverifiable and unobservable than usual in the following sense. In the case of the CDE above, we can at least imagine running a randomized experiment in which we randomly assign treatment and mediator. In that case, both assumptions would be met and the CDE identified. To then move to the observational setting, we just have to reason about why it is plausible (if it is) that $M_i(t)$ is 'like random' conditional on $T_i = t, X_i = x$. In the NIE/NDE setting, there is no ideal experimental version: if you were to randomize M, it wouldn't be a natural effect anymore! Relatedly, unlike $Y_i(t, m)$, the $Y_i(t, M_i(t'))$ for $t \neq t'$ type quantities are fundamentally unobservable under any treatment assignment! Some view this as a philosophical leap too far.

   See Module 9 review questions 2 for an alternative identifying assumption that avoids using cross-world counter-factuals.

Running Example: **How can the cross-worlds counterfactual assumption be violated?** for simplicity, ignore conditioning on $X_i = x$. Suppose we consider units who received health insurance ($T_i = 1$). The cross-world counterfactual assumption says that whether or not these units receive surgery $M_i(t)$ in this observed world is independent of what their outcomes *would have been* under surgery $m = 1$ or no surgery $m = 0$ if they had not received health insurance. Now suppose there is actually a preventative care variable playing the role of Z in Figure 27. Suppose getting preventative care can be affected by treatment and affects whether people get surgery (M) and their outcome (Y). For simplicity, suppose people can only get the preventative care if they have health insurance so that only $Z_i(0) = 0, Z_i(1) = 0$ (never-takers) and $Z_i(0) = 0, Z_i(1) = 1$ (compliers) preventative care strata exist. Suppose preventative care makes people less likely to get surgery and also separately causes better health outcomes. Suppose also that the compliers are systematically healthier people than the never-takers[52] so that even without health insurance, and regardless of surgery, they have better outcomes. By randomization of treatment, we should have both compliers and never-takers in our $T_i = 1$ group. However, we now have a dynamic where:

- Learning someone in the treatment group got surgery $M_i(1) = 1$ makes it more likely they are never-takers of preventative care, which means their $Y_i(0, m)$ (cross world) tend to be low.

[52] e.g., they might be people more concerned about having a healthy lifestyle

- Learning someone in the treatment group did not get surgery makes it more likely they are compliers, which means their $Y_i(0, m)$ (cross world) tend to be high.

This would violate the cross-world assumption. In short, the key problem is that other mediators with a relationship to $M_i(t)$ hurt our ability to identify the role of just the one of interest. Intuitively, if preventative care is also playing a role, we might observe people with surgery having worse outcomes than people without surgery but only because those without surgery got preventative care. It might be that really, among those with $Z_i = z$ fixed, surgery would lead to better Y than non-surgery, but when the focus is *natural* effects, we do not get to fix post-treatment outcomes like that.

Do we need to worry about Z_i? as the above example illustrates, there are realistic scenarios where a Z_i can occur. One strategy is to focus on mediators M_i which happen very soon after treatment so that there is less time for there to be other mediators which also affect M_i. The problem with surgery

9.3.1 Identification

Let's now look how these assumptions allow us to identify the NIE and NDE. We will consider the case of $Y_i(t, M_i(t))$ and $Y_i(t, M_i(1-t))$ separately. The first is easily identified without any cross-world counterfactual requirement.

$$\begin{aligned}
\mathbb{E}(Y_i(t, M_i(t))) &= \mathbb{E}(\ \mathbb{E}(Y_i(t, M_i(t))|X_i)) && \text{(Law of Total Exp)} \\
&= \mathbb{E}(\ \mathbb{E}(Y_i(t, M_i(t))|T_i = t, X_i)) && \text{(Assumption 1)} \\
&= \mathbb{E}(\ \mathbb{E}(Y_i|T_i = t, X_i)) && \text{(Consistency)} \\
&= \sum_x \mathbb{E}(Y_i|T_i = t, X_i = x) \Pr(X_i = x) && \text{(Write out expectation)} \\
&= \sum_x \sum_m \mathbb{E}(Y_i|M_i = m, T_i = t, X_i = x) \Pr(M_i = m|T_i = t, X_i = x) \Pr(X_i = x) && \text{(Law of Total Probability)}
\end{aligned}$$

Writing it out as in the last line will be useful below. The right quantity is where we need the **cross world counterfactual.**

$$\begin{aligned}
\mathbb{E}(Y_i(t, M_i(1-t)) &= \sum_x \mathbb{E}(Y_i(t, M_i(1-t))|X_i = x)\Pr(X_i = x) && \text{(Law of Tot. Exp.)} \\
&= \sum_x \mathbb{E}(Y_i(t, M_i(1-t))|T_i = t, X_i = x) \Pr(X_i = x) && \text{(Assumption 1)} \\
&= \sum_x \sum_m \mathbb{E}(Y_i(t, M_i(1-t)) \mid M_i(1-t) = m, T_i = t, X_i = x) \\
&\qquad \Pr(M_i(1-t) = m|X_i = x, T_i = t) \Pr(X_i = x) && \text{(Law of Tot. Exp.; no A2 yet)} \\
&= \sum_x \sum_m \mathbb{E}(Y_i(t, m) \mid T_i = t, X_i = x) \\
&\qquad \Pr(M_i(1-t) = m|X_i = x, T_i = t) \Pr(X_i = x) && \text{(Cross-World Counterfactual)} \\
&= \sum_x \sum_m \mathbb{E}(Y_i(t, m) \mid M_i(t) = m, T_i = t, X_i = x) \\
&\qquad \Pr(M_i(1-t) = m|X_i = x, T_i = t) \Pr(X_i = x) && \text{(Assumption 2 with } t = t') \\
&= \sum_x \sum_m \mathbb{E}(Y_i(t, m) \mid M_i(t) = m, T_i = t, X_i = x) \\
&\qquad \Pr(M_i(1-t) = m|X_i = x, T_i = 1-t) \Pr(X_i = x) && \text{(Assumption 1)} \\
&= \sum_x \sum_m \mathbb{E}(Y_i \mid M_i = m, T_i = t, X_i = x) \\
&\qquad \Pr(M_i = m|X_i = x, T_i = 1-t) \Pr(X_i = x) && \text{(Consistency)}
\end{aligned}$$

Note that the summations can all be replaced with integrals if x or m is continuous. Note also that the above is the NIE if $t = 1$ and $-$NIE if $t = 0$.

Note: see also Module 9 Review question 2a for this proof.

Identification of NIE: for $\bar{\delta}(t) = \mathbb{E}(Y_i(t, M_i(1))) - \mathbb{E}(Y_i(t, M_i(0)))$, we can combine the above as follows:

$$\bar{\delta}(t) = \mathbb{E}(Y_i(t, M_i(t)) - \mathbb{E}(Y_i(t, M_i(1-t)) = \sum_{x,m} \mathbb{E}(Y_i|M_i = m, T_i = t, X_i = x) * \\ [\Pr(M_i = m|T_i = t, X_i = x) - \Pr(M_i = m|T_i = 1-t, X_i = x)] * \Pr(X_i = x)$$

Identification of NDE: we have $\bar{\zeta}(t) = \mathbb{E}(Y_i(1, M_i(t)) - Y_i(0, M_i(t)))$ and using the results, get:

$$\bar{\zeta}(t) = \mathbb{E}(Y_i(1, M_i(1)) - Y_i(0, M_i(1))) = \sum_{x,m} [\mathbb{E}(Y_i|M_i = m, T_i = 1, X_i = x) - \mathbb{E}(Y_i|M_i = m, T_i = 0)] * \\ \Pr(M_i = m|T_i = t, X_i = x) * \Pr(X_i = x)$$

Identification of Total Effect: recall that the total effect $\mathbb{E}(Y_i(1, M_i(1)) - Y_i(0, M_i(0))$ can be decomposed as in Section 9.1.4. The total effect is easily identifiable using the just the first derivation above as:

$$\mathbb{E}(Y_i(1, M_i(1)) - Y_i(0, M_i(0)) = \sum_x [\mathbb{E}(Y_i|T_i = 1, X_i = x) - \mathbb{E}(Y_i|T_i = 0, X_i = x)] * \Pr(X_i = x)$$

9.3.2 Estimation of NIE and NDE

We are now left with the task of estimating each of the observational quantities in the identification equalities above. We need to estimate:

1. $\Pr(X_i = x)$
2. $\Pr(M_i = m|X_i = x, T_i = t)$
3. $\mathbb{E}(Y_i|M_i = m, T_i = t, X_i = x)$

If everything is discrete and not high dimensional, these may all be estimable via sample proportions. Otherwise, we will require some additional modeling assumptions. The different general estimation approaches described in Section 9.2.2 apply here, too:

1. **Plug in models:** If our models are tractable to write down in closed form, we might simply plug them into our identification equalities for the NDE and NIE. Again, this is not necessarily an optimal approach.
2. **Prediction-based**:
   For the NIE, we can use our model of $\Pr(M_i = m|X_i = x, T_i = t)$ for $t = 0$ and $t = 1$ to predict $M_i(1)$ and $M_i(0)$ and plug these in into the outcome model in turn:

   $$\hat{\delta}(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(Y_i|\widehat{M_i(1)}, T_i = t, X_i) - \mathbb{E}(Y_i|\widehat{M_i(0)}, T_i = t, X_i)$$

   One way to think of
   For the NDE, we predict $M_i(t)$ only once and plug it into both models as below

   $$\hat{\zeta}(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(Y_i|\widehat{M_i(t)}, T_i = 1, X_i) - \mathbb{E}(Y_i|\widehat{M_i(t)}, T_i = 0, X_i)$$

   Notice that in the NIE, $T_i = t$ reflects the value at which we are holding the treatment constant while the M_i value plugged in gets to vary while for the NDE, we look at $T_i = 1, T_i = 0$ and fix the mediator. This is in line with the meaning of these two estimands.
3. **Leverage decomposition**: We could also estimate the total effect easily using just an outcome model $\mathbb{E}(Y_i|T_i = t, X_i)$ and subtract an estimate of the NDE for $1 - t$ to obtain an estimate of the NIE for t or vice versa (see Section 9.1.4)

9.3.3 Estimation under linearity assumption

Suppose that Y_i and M_i are continuous-valued and we are willing to assume they have a linear relationship to X_i. The DAG in Figure 28 then represents the following Linear Structural Equations (Section 4.1 and 7.6.2)

$$\begin{aligned} Y_i &= \alpha_1 + \beta_1 T_i + \lambda_1^T X_i + \epsilon_{1i} \\ M_i &= \alpha_2 + \beta_2 T_i + \lambda_2^T X_i + \epsilon_{2i} \\ Y_i &= \alpha_3 + \beta_3 T_i + \gamma M_i + \lambda_3^T X_i + \epsilon_{3i} \end{aligned}$$

The first regression is actually redundant but β_1 represents the **total effect** of T_i on Y_i (in terms of the DAG, imagine collapsing M_i into the arrow $T \to Y$, as we were implicitly doing in every module so far). The **direct effect** of treatment is equal to β_3 in the third equation while the indirect effect is $\gamma\beta_2$, the product of the effect of changing T_i from 0 to 1 and the effect of changing M_i. To see this, we can substitute the second equation into the third:

$$\begin{aligned} Y_i &= \alpha_3 + \beta_3 T_i + \gamma(\alpha_2 + \beta_2 T_i + \lambda_2^T X_i + \epsilon_{2i}) + \lambda_3^T X_i + \epsilon_{3i} \\ &= \alpha_3 + (\beta_3 + \gamma\beta_2) T_i + (\gamma\lambda_2 + \lambda_3)^T X_i + (\gamma\epsilon_{2i} + \epsilon_{3i}) \end{aligned}$$

In terms of the first equation above, we then have $\alpha_3 = \alpha_1$, $\beta_1 = \beta_3 + \gamma\beta_2$ (this is the total effect decomposition), $\lambda_1 = \gamma\lambda_2 + \lambda_3$ and $\epsilon_{1i} = \gamma\epsilon_{2i} + \epsilon_{3i}$. This also implies that the indirect effect may be calculated as $\beta_1 - \beta_3$. Overall, if we are willing to assume linearity, we can identify direct and indirect effects by fitting either the first two regressions or the second two regressions above. Implicitly, the above still assumes unconfoundedness (of treatment + no post-treatment confounders) via the regression exogeneity assumption.

There are further extensions of this, including models which allow for the effect of the mediator to change depending on the level of treatment by adding a mediator-treatment interaction (see (Slide 11).

10 Module 10: Panel Data: Difference in Differences, Lagged Outcome, Fixed Effects Models, and Synthetic Control

Big Picture: in previous modules, we confronted the dangers of confounders– ways in which the treated and control group might be systematically different for reasons other than treatment, even after all our best efforts to condition and match and weight our way to similarity. But what if we have *repeated measurements* of treated and control units and what if those units, though systematically different, are evolving over time in a similar way in some baseline period before anyone is treated? If one unit is then treated and nothing else systematically changes between the two, perhaps we can treat the subsequent *trend* in the untreated unit as an indication of what trend would have happened to the treated unit if untreated! In this module, we leverage repeated measurements (also known as panel data, longitudinal data) to approach the counterfactual world by this route.

10.1 Difference in Differences (DiD)

DiD is a very popular methodology because, as described above, it allows for the presence of **time-invariant confounders**. That is, the treated and control group are allowed to be systematically different as long as those differences are stable over time so that it is reasonable to assume that the only possible disrupter of the pre-treatment correlation between treated unit and control unit outcomes over time is the treatment itself. Intuitively, if we observe that after being treated, the treated units' outcomes diverge from what we *would have predicted them to be* based on pre-treatment trends, then that would seem to reflect an effect of treatment. The idea is commonly illustrated as follows:

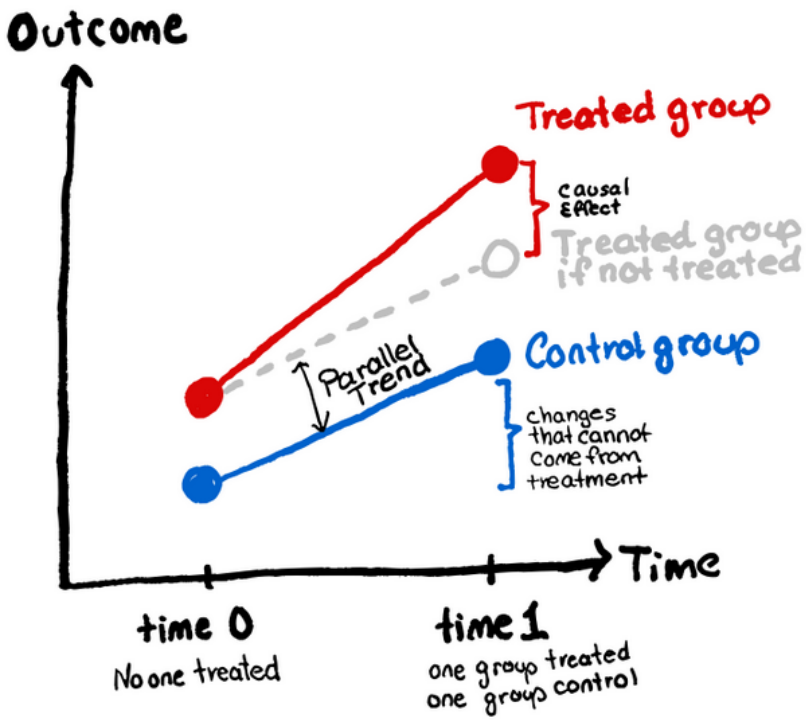


Figure 29: Basic DiD Diagram: Illustration of Parallel Trends assumption.

Figure 29 shows average potential outcome trends for two groups of units. Even at baseline time 0, these groups have different outcomes. Sometime between time 0 and time 1, one group is treated and unfortunately, we then cannot observe its outcomes if it had not been treated (gray line). However, we do observe outcomes at time 1 for untreated units in the other group. If the change over time for that group, captures how the outcomes of the treated group would have changed even if not treated (parallel trends), then we can subtract that difference for the control group from the observed average treated group outcome at time 1 to estimate the average treated group outcome under control at time 1. We will formalize all of this below.

Caution: although many of the diagrams below contain lines meant to make it easier to visualize the hypothesized relationships, in the 2×2 DiD case (as this is sometimes called), we only ever observe data at discrete time points 0 and time 1 and only make assumptions about how potential outcomes at these time points relate.

Example: As on Slide 3, consider the example from Card and Krueger (1993). This study evaluated the effect of increasing the minimum wage on employment by comparing the US State of New Jersey, where minimum wage increased, to its neighbor, Pennsylvania, where it did not. The idea is that although there might be some systematic differences between NJ and PA, the two being neighbors means that, absent any treatment (wage increase) their general employment trends would plausibly be similar, rising and falling over time in the same general way due to various regional and macroeconomic factors. If we assume that the trend in employment in PA is the trend NJ would have had without treatment, then DiD allows us to use this to back out what NJ's outcome under non-treatment would have been. As shown in Figure 30, in this case there was a quite dramatic dynamic where in NJ, the employment actually went up even though baseline employment was trending down!

As always in causal inference, we want to think carefully about what could go wrong here to make us think there is a causal effect when there is not one. There are at least two dynamics to worry about: 1. If PA and NJ were actually very decoupled and dissimilar, the parallel trends idea might be implausible. For example, imagine we replaced PA with Puerto Rico. It seems a lot less plausible that employment trends in Puerto Rico would reflect what would have happened in NJ if the minimum wage had not been increased. 2. Even if NJ and PA are similar, there is the risk of something else happening in NJ but not in PA between $t = 0$ and $t = 1$ so that the observed spike in employment is actually attributable to that variable and not to the minimum wage increase. Hence, to make their causal argument plausible, the researchers have to argue that NJ and PA are similar enough to expect parallel trends and that no other NJ-only or PA-only shock happened in the relevant time period.

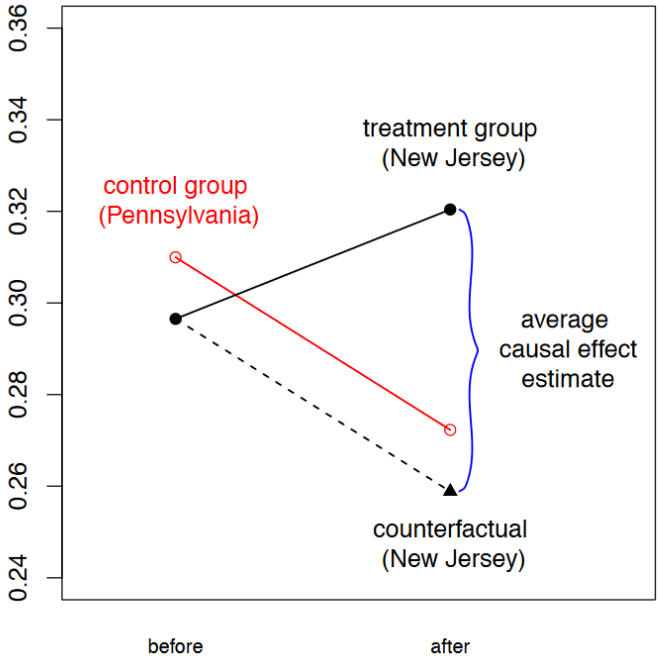


Figure 30: Example from Card and Krueger (1993), plot copied from Slide 4. The black dotted line represents the parallel trends assumption.

The DiD Advantage: Leveraging Multiple Comparison Types

The intuition described above reflects two important kinds of comparisons, only one of which we have considered so far in the course. All of the previous modules focused on **cross-sectional** causal comparisons, which are between groups of distinct units, some treated, and some not. For example, when we imagined an experiment where some children are tutored and some children are not, this was cross-sectional. Usually, we would hope that the experiment observes children treated and not treated within the same time window so as to avoid confounding due to factors that vary over time. However, even if there is no confounding, a cross-sectional study can have the problem that there is a lot of heterogeneity of outcomes within treatment and outcome groups due to factors other than the treatment so that it is hard to detect causal effects.

In contrast, **temporal** causal comparisons make before/after type arguments for repeated observations of the same unit. For example, if children receive a pre-test, then are tutored, and then receive a post-test, we might want to point to higher post-test scores as evidence of tutoring effectiveness. The problem with such a comparison is that it does not tell us how the children would have performed on the post-test if they had not been tutored and *that* comparison is the actual causal effect – there is no control group! But the appeal of such comparisons is that by comparing outcomes for the same child, we potentially control for many other sources of variation.

Figure 31 illustrates the limitations of each approach by clarifying what it is assuming about potential outcomes. In the left plot, imagine we have a group which we observe under control at time t and under treatment at time $t+1$. If we do this before/after comparison and call it causal, then we are assuming that $\mathbb{E}(Y_{i,time=0}(treat=0)) = \mathbb{E}(Y_{i,time=1}(treat=0))$. This is represented by the blue line in the plot. The right plot shows what happens if we only do a cross-sectional comparison at time $t+1$. In this case, the blue dotted line represents the assumption that the observed control group at time $t+1$ represents how the treatment group would have evolved relative to its value time t. Hence the left plot fails to account for change over time while the right plot fails to account for systematic difference between the treated and control groups. **Difference in Differences is powerful because it leverages both these types of comparisons** and thereby avoids the limitations suggested in the plots.

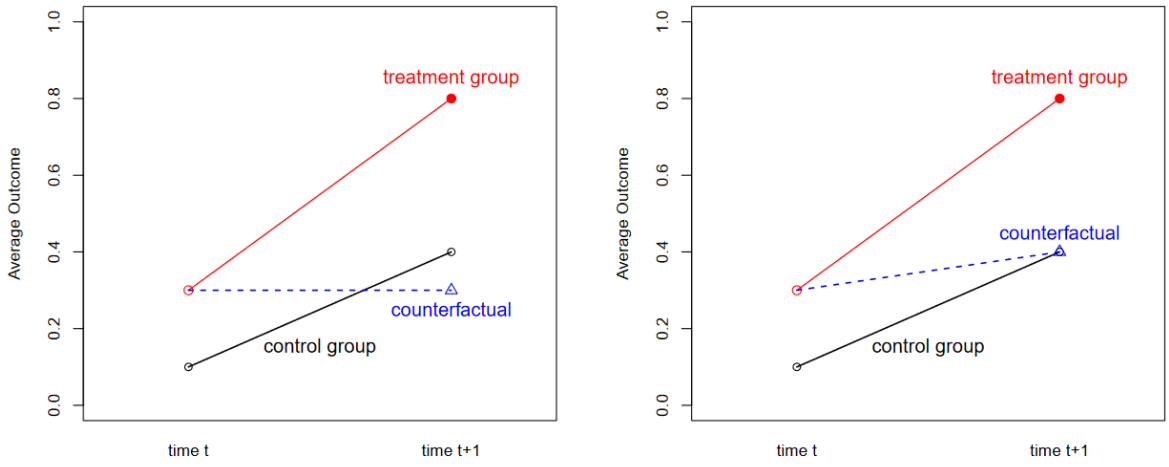


Figure 31: Figure from Slide 2.

10.1.1 Set-up, Estimand, Identification, Estimation

Because it is natural to let t represent time, we shift notation a bit for difference in differences and let g be the index for treatment. The basic DiD set-up is as follows:

- At time 0, no units are treated
- At time 1, some units are treated and others are not
- $Z_{it} = tG_i$ indicates whether unit i at time t received treatment. E.g., a unit in group $G_i = 1$ has $Z_{i0} = 0$ and $Z_{i1} = 1$.
- $G_i \in \{0, 1\}$ is a time-invariant indicator of *membership* in treatment or control group. It is 1 for units that are treated at time 1 and 0 for the rest. [53]
- Every unit has four potential outcomes $Y_{i,time}(treatment)$: $Y_{i0}(0), Y_{i0}(1), Y_{i1}(0), Y_{i1}(1)$ (See Figure 32). $Y_{i0}(1)$, the outcome under treatment at time 0, hypothetically exists but is never observed in the basic DiD set-up. For each unit, $Y_{i0}(0)$ and one of the other two is observed.
- Assuming consistency, the observed outcomes are $Y_{it} = Y_{it}(Z_{it})$

Using this notation, our **estimand** is an average treatment effect on the treated units at time 1.

$$\tau := \mathbb{E}(Y_{i1}(1) - Y_{i1}(0)|G_i = 1)$$

Data on time 0 and on the control units help identify τ but we do not estimate anything directly about those contexts. For identification, the one crucial assumption is the **Parallel Trends Assumption**:

$$\mathbb{E}[Y_{i1}(0) - Y_{i0}(0)|G_i = 1] = \mathbb{E}[Y_{i1}(0) - Y_{i0}(0)|G_i = 0]$$

Notice that every potential outcome in this assumption is under control. Because units with $G_i = 1$ are treated at time 1, $Y_{i1}(0)$ is the one unobservable quantity involved. The key assumption is that, on average, the trend for the control group and the trend for the treated group if it had not been treated are the same. This is diagrammed for two units i, j in Figure 32, though in practice we only require it to hold on average. That said, in some cases such as the NJ vs PA comparison example above, a DiD study will truly have only one unit in each group and then the parallel trend assumption must hold at the individual level.

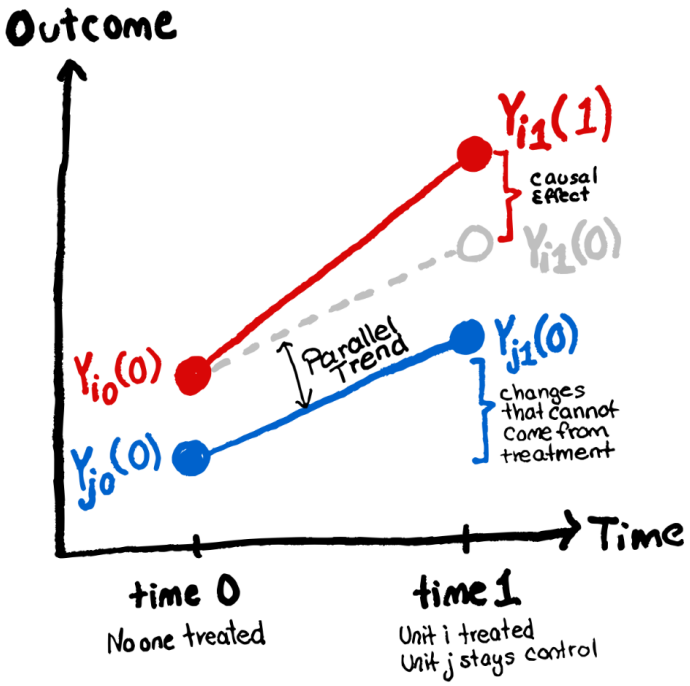


Figure 32: Basic DiD Diagram: Illustration of Parallel Trends assumption.

[53] Note: technically this notation is redundant. We could just condition everywhere on $(Z_{i0}, Z_{i1}) = (0, 0)$ and $(Z_{i0}, Z_{i1}) = (0, 1)$. But the G_i makes this much simpler to write down and deal with.

Note that the parallel trends assumption is not a **testable** assumption. No data can tell us whether it truly holds. However, if we have multiple pre-treatment time period measurements, it is possible to further support its plausibility by doing **placebo tests** on pre-treatment time periods. That is, if in a longer pre-treatment period, the two groups have similar trends, that does support (but not prove) that absent any other shocks between $t = 0$ and $t = 1$, parallel trends would hold.

Given the parallel trends assumption, the **identification** argument can be broken into two parts. First, by consistency, we immediately can identify the average outcome under treatment for the treated group at time 1.

$$\mathbb{E}(Y_{i1}(1)|G_i = 1) = E(Y_{i1}|G_i = 1)$$

For the average outcome under control of the treatment group at time 1, we solve for the quantity of interest in the parallel trends assumption to obtain:

$$\begin{aligned}\mathbb{E}(Y_{i1}(0)|G_i = 1) &= \mathbb{E}[Y_{i1}(0) - Y_{i0}(0)|G_i = 0] + \mathbb{E}[Y_{i0}(0)|G_i = 1] \quad \text{(parallel trends)} \\ &= \mathbb{E}[Y_{i1} - Y_{i0}|G_i = 0] + \mathbb{E}[Y_{i0}|G_i = 1] \quad \text{(consistency)}\end{aligned}$$

Putting these together:

$$\begin{aligned}\tau_{DiD} &= \mathbb{E}(Y_{i1}(1)|G_i = 1) - \mathbb{E}(Y_{i1}(0)|G_i = 1) \\ &= [\mathbb{E}(Y_{i1}|G_i = 1) - \mathbb{E}(Y_{i0}|G_i = 1)] \; - \; [\mathbb{E}(Y_{i1}|G_i = 0) - \mathbb{E}(Y_{i0}|G_i = 0)] \\ &= \mathbb{E}(Y_{i1} - Y_{i0}|G_i = 1) \; - \; \mathbb{E}(Y_{i1} - Y_{i0}|G_i = 0)\end{aligned}$$

The identification equality above is entirely **non-parametric**, and one immediate way to obtain an estimator is simply to replace each of the expectations with its relevant sample quantity. Letting n_1 and n_0 be the number of treated and control units, we have the difference in means estimator

$$\hat{\tau}_{DiD} = \frac{1}{n_1}\sum_{i=1}^{n} G_i(Y_{i1} - Y_{i0}) - \frac{1}{n_0}\sum_{i=1}^{n}(1 - G_i)(Y_{i1} - Y_{i0}) \tag{25}$$

Notice this is exactly a difference in the differences at time 1 vs 0 for each group.

Note: the above estimator assumes a **panel data** set-up (the same exact units appear at each time 0 and 1) but we can also use Difference in Differences for settings where we have a separate random sample from the treatment and control group at each time point. We would then take the difference in the time 1 vs time 0 differences for each group. This comes from writing the identification equality as:

$$\tau_{DiD} = [\mathbb{E}(Y_{i1}|G_i = 1) - \mathbb{E}(Y_{i0}|G_i = 1)] \; - \; [\mathbb{E}(Y_{i1}|G_i = 0) - \mathbb{E}(Y_{i0}|G_i = 0)]$$

Notice that if G_i were randomized, this would reduce to $\mathbb{E}(Y_{i1}(1)) - \mathbb{E}(Y_{i1}(0))$, exactly the average treatment effect over all observations at a single time 1 as we studied for example in Module 3.

10.1.2 Connecting DiD and Regression

In the basic 2 × 2 DiD case described so far, this estimator can be connected to a particular regression, much like we did in Module 4 and Module 5 where it was desirable to give practitioners a regression type model that they could fit using standard software. We first reformulate the theoretical set-up in terms of the following **two-way fixed effects** structural linear model:[54]

$$Y_{it}(z) = \alpha_i + \beta t + \tau z + \epsilon_{it} \tag{26}$$

where

- α_i is a unit-level fixed effect. This does not just collapse into the error term because we have repeated measurements so α_i is constant over time.
- β is the time fixed effect. This effect of time is constant across units.
- τ is a fixed effect of treatment
- ϵ_{it} is random error of unit i at time t. This is the remaining individual-observation-level variation. As in Section 4.1.1, we could also have $\epsilon_{it}(z)$
- Without further assumption, $\mathbb{E}(\epsilon_{it}) = 0$ for the same reason discussed in Section 4.1.1.
- Because every variable is treated categorically, this model imposes no linearity assumption

We then have $\mathbb{E}(Y_{i0}(0)) = \alpha_i$, $\mathbb{E}(Y_{i1}(0)) = \alpha_i + \beta$ and $\mathbb{E}(Y_{i1}(1)) = \alpha_i + \beta + \tau$ and hence

$$\begin{aligned}\mathbb{E}(Y_{i1}(1) - Y_{i1}(0)) &= (\alpha_i + \beta + \tau - \alpha_i - \beta) = \tau \\ \mathbb{E}(Y_{i0}(1) - Y_{i0}(0)) &= (\alpha_i + \tau - \alpha_i) = \tau\end{aligned}$$

Technically, this model poses that the effect of treatment is constant over time, but since in DiD, we only have treated observations at time 1, we are focused on the first $Y_{i1}(1) - Y_{i1}(0)$ case. Note, however, that this is not the parallel trends assumption. The **parallel trends** assumption can be formulated as

$$\mathbb{E}[Y_{i1}(0) - Y_{i0}(0)|G_i = g) = \beta \quad \text{for} \quad g = 0, 1$$

Equivalently, this imposes a constraint on the error terms:

$$\begin{aligned}\mathbb{E}(\alpha_i + \beta + \epsilon_{i1} - \alpha_i - \epsilon_{i0}|G_i = g) = \mathbb{E}(\epsilon_{i1} + \beta - \epsilon_{i0}|G_i = g) &= \beta \\ \mathbb{E}(\epsilon_{i1} - \epsilon_{i0}|G_i = g) &= 0 \qquad \text{for } g = 0, 1\end{aligned}$$

Notice that this has a similar form to the exogeneity assumption discussed in Module 4 if we think of the outcome as being $Y_{i1} - Y_{i0}$. It says that on average, the error terms should not have a systematic trend over time – i.e., no time-varying confounders. On the other hand, time invariant confounders are allowed. These would show up in the α_i values varying systematically with the value of z, which the model does not preclude. If instead of Equation 26, the true model has a **time-treatment interaction**

$$Y_{it}(z) = \alpha_i + \beta t + \tau z + \gamma t z + \eta_{it}$$

this would violate parallel trends as formulated above in terms of ϵ_{it}. That is, from the perspective of our previous model, we now have an error term $\epsilon_{it} = \eta_{it} + \gamma t z$ and hence,[55]

$$\begin{aligned}\mathbb{E}(\epsilon_{i1} - \epsilon_{i0}|G_i = 1) &= \gamma + \mathbb{E}(\eta_{i1} - \eta_{i0}|G_i = 1) \\ \mathbb{E}(\epsilon_{i1} - \epsilon_{i0}|G_i = 0) &= \mathbb{E}(\eta_{i1} - \eta_{i0}|G_i = 0)\end{aligned}$$

Here, even if $\mathbb{E}(\eta_{i1} - \eta_{i0}|G_i = g) = 0$, we are left with a γ capturing the tendency for the error terms to vary systematically over time.

[54] Terminology: this model has fixed effects for unit, time, and treatment as opposed to random effects. The 'two' refers to there being a unit fixed effect and a time fixed effect (the τ treatment effect is not part of the count for some reason...). 'Structural' reflects that this is a causal model.

[55] Here I am implicitly using the fact that at time 0 a $G_i = 1$ unit has $tz = 0 * 0 = 0$ and at time 1, $tz = 1 * 1 = 1$

Connecting regression estimator to DiD estimator:

None of the discussion above proves that fitting the non-interaction regression with our observed data gives exactly the $\hat{\tau}_{DiD}$ estimator. We've really just changed notation and written down what parallel trends means in the new notation. To think about fitting the regression, we can imagine formatting our dataset in a single table with a column for time T_i and treatment Z_{iT_i} and with all of our $Y_{iT_i} = Y_{iT_i}(Z_i)$'s in a column as well. The regression then minimizes:

$$\underset{\alpha_1,\ldots,\alpha_n,\beta,\tau}{\arg\min} \sum_{i=1}^{n}\sum_{t=0}^{1}(Y_{it} - \alpha_i - \beta t - \tau Z_{it})^2$$

To prove this is equivalent to the DiD estimator, we use the **Frisch-Waugh-Lovell-Theorem**, which tells us that it is algebraically equivalent to first regress everything on an indicator for the unit to get rid of the unit specific effects.[56] We then regress the residuals on each other. Because we are regressing on categorical covariates, it is easy to write down what form these residuals take:

- $\tilde{Y}_{it} = Y_{it} - \frac{Y_{i1}+Y_{i0}}{2}$
- $\tilde{Z}_{it} = Z_{it} - \frac{Z_{i1}+Z_{i0}}{2} = \begin{cases} \frac{1}{2} & \text{if } t=1, G_i = 1 \\ -\frac{1}{2} & \text{if } t=0, G_i = 1 \\ 0 & \text{if } G_i = 0 \end{cases}$
- $\tilde{t} = \begin{cases} 1 - \frac{1}{2} = \frac{1}{2} & \text{if } t = 1 \\ 0 - \frac{1}{2} = -\frac{1}{2} & \text{if } t = 0 \end{cases}$ (all units are observed at time 0 and 1 so their average time is $\frac{1}{2}$)

The second-stage FWL regression is then:

$$\begin{aligned}
\underset{\beta,\tau}{\arg\min} \sum_{i=1}^{n}\sum_{t=0}^{1}(\tilde{Y}_{it} - \beta\tilde{t} - \tau\tilde{Z}_{it})^2 &= \underset{\beta,\tau}{\arg\min} \sum_{i=1}^{n}(\tilde{Y}_{i1} - \beta\frac{1}{2} - \tau\tilde{Z}_{i1})^2 + \sum_{i=1}^{n}(\tilde{Y}_{i0} + \beta\frac{1}{2} - \tau\tilde{Z}_{i0})^2 \quad \text{(break into t=0,t=1 cases)} \\
&= \underset{\beta,\tau}{\arg\min} \sum_{i=1}^{n}(Y_{i1} - \frac{Y_{i1}+Y_{i0}}{2} - \beta\frac{1}{2} - \frac{1}{2}\tau G_i)^2 + \sum_{i=1}^{n}(Y_{i0} - \frac{Y_{i1}+Y_{i0}}{2} + \beta\frac{1}{2} + \frac{1}{2}\tau G_i)^2 \\
&= \underset{\beta,\tau}{\arg\min} \sum_{i=1}^{n}(\frac{Y_{i1}-Y_{i0}}{2} - \beta\frac{1}{2} - \frac{1}{2}\tau G_i)^2 + \sum_{i=1}^{n}(-\frac{Y_{i1}-Y_{i0}}{2} + \beta\frac{1}{2} + \frac{1}{2}\tau G_i)^2 \\
&= \underset{\beta,\tau}{\arg\min} \sum_{i=1}^{n}(\frac{Y_{i1}-Y_{i0}}{2} - \beta\frac{1}{2} - \frac{1}{2}\tau G_i)^2 + \sum_{i=1}^{n}(\frac{Y_{i1}-Y_{i0}}{2} - \beta\frac{1}{2} - \frac{1}{2}\tau G_i)^2 \quad \text{(pull out -1 on right)} \\
&= 2 * \underset{\beta,\tau}{\arg\min} \sum_{i=1}^{n}(\frac{Y_{i1}-Y_{i0}}{2} - \beta\frac{1}{2} - \frac{1}{2}\tau G_i)^2 \\
&= \underset{\beta,\tau}{\arg\min} \sum_{i=1}^{n}(Y_{i1} - Y_{i0} - \beta - \tau G_i)^2
\end{aligned}$$

This last line is exactly a regression of $Y_{i1} - Y_{i0}$ on a binary random variable G_i. By the same equivalence as in Module 4 (Section 4.2), the OLS estimate $\hat{\tau}_{reg}$ is the difference in means for $G_i = 1$ and $G_i = 0$, which is exactly $\hat{\tau}_{DiD}$ (equation 25).

Caution: this equivalence does NOT hold in more general DiD set-ups. One cannot simply run the fixed effects model and call it difference in differences in general.

[56] You can imagine dummy-encoding all the unit-level fixed effects by defining variable I_{ij} for each unit which is 1 only if $i = j$ and 0 otherwise. Then each unit i has an associated vector $(I_{i1}, ..., I_{in})$ and we add n terms $\alpha_j I_{ij}$ in the regression above. We are first regressing on these indicators.

10.1.3 DiD with Covariates

Many of the ideas developed throughout this course can be adapted to DiD. For example, we can do DiD while adjusting for pre-treatment covariates. In this case, we make a conditional parallel trends assumption

$$\mathbb{E}[Y_{i1}(0) - Y_{i0}(0)|X_i = x, G_i = 1] - \mathbb{E}[Y_{i1}(0) - Y_{i0}(0)|X_i = x, G_i = 0] \quad \forall x$$

The intuition is the same as in Module 7. If it is implausible that parallel trends holds unconditionally, it might be more plausible that the necessary assumption holds conditionally. One way to think about what we are doing here is that we control for observed confounders via conditioning and for unobserved (time-invariant) confounders via parallel trends. Note however that the conditional parallel trends assumption neither implies nor is implied by the unconditional parallel trends assumption – it is not always the case that conditioning makes the assumption more plausible, so it is important to be careful about what you condition on. Figure 33 gives an illustrative example: Imagine we have groups 1 and 2 such that:

- Group 1 has a positive treatment and baseline trend
- Group 2 has only a slight positive treatment trend and a negative baseline trend
- Group 1 has many more treated units than control and Group 0 has many more control units than treated

Suppose in this scenario it is plausible that parallel trends holds within each group. As shown in the diagram, when we pool them, averaging dynamics can lead to violations of parallel trends. Essentially, the pooled control group is not similar enough to the pooled treatment group because of a hidden confounder X = group. Group 1 and Group 2 are changing over time in different ways and because the pooled control and pooled treatment group are made up of these groups in different proportions, this makes the pooled control group change differently over time from how the pooled treatment group would have changed over time under control.

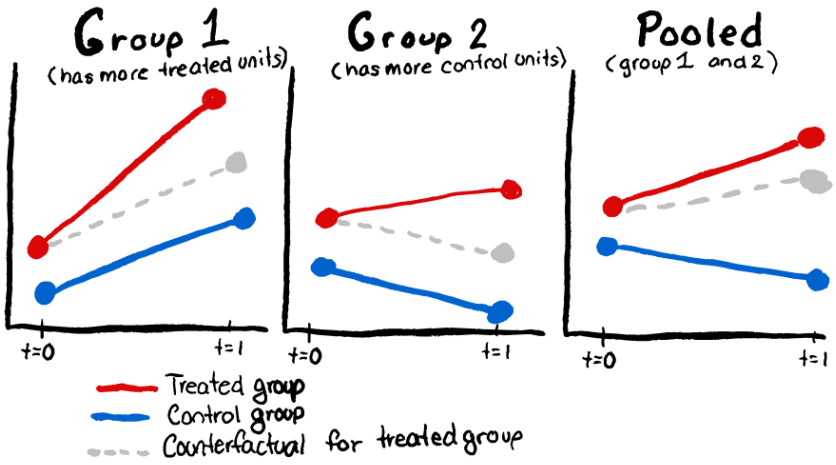


Figure 33: Illustration of parallel trends holding conditionally but not unconditionally

The identification and one possible estimation strategy for the overall estimand τ in this set-up is essentially the same as the previous only now with conditioning on X_i via a law of total expectation step.

$$\begin{aligned}
\tau_{DiD} &= \mathbb{E}(Y_{i1}(1)|G_i = 1) - \mathbb{E}(Y_{i1}(0)|G_i = 1) \\
&= \mathbb{E}(\mathbb{E}(Y_{i1}(1)|X_i, G_i = 1) - \mathbb{E}(\mathbb{E}(Y_{i1}(0)|X_i, G_i = 1)) \\
&= [\mathbb{E}(\mathbb{E}(Y_{i1}|X_i, G_i = 1)) - \mathbb{E}(\mathbb{E}(Y_{i0}(0)|X_i, G_i = 1))] \; - \; [\mathbb{E}(\mathbb{E}(Y_{i1}(0)|X_i, G_i = 0)) - \mathbb{E}(\mathbb{E}(Y_{i0}(0)|X_i, G_i = 0))] \\
&= \mathbb{E}(\mathbb{E}(Y_{i1} - Y_{i0}|X_i, G_i = 1) \; - \; \mathbb{E}(\mathbb{E}(Y_{i1} - Y_{i0}|X_i, G_i = 0))
\end{aligned}$$

If we have panel data, we could estimate these by regressing $Y_{i1} - Y_{i0}$ on X_i and $G_i = g$ for $g = 0, 1$. If X_i is continuous, this will involve further modelling assumptions such as linearity. If we do not have panel data and instead have a separate random sample of units at each time period and condition, we can fit four regressions, one for each $\mathbb{E}(Y_{it}|G_i = g, X_i)$ and combine them as in the identification equality. As noted on Slide 9, there are also matching and weighting based approaches one can use instead of doing outcome regressions (see Abadie (2005)). There are also doubly robust estimators that combine the two with the usual result that the estimator is consistent if one of the two models is correct (see Sant'Anna and Zhao (2020)).

10.1.4 Miscellaneous Notes on DiD

- **Relationship of DiD to RD?** All this talk of trends and some kind of continuity of the trend might remind you of Regression Discontinuity from Module 6. There are some similarities here, but while RD involves a running variable with a cut-off such that everyone on one side gets treated and everyone on the other does not, in DiD, the key juncture is a point where some are treated and some are not. On one side of the cut-off, *no one is treated* while on the other, some are. That said, there are various parallels between the two. For example, RD and DiD are the two identification strategies we have seen in this course that do not rely on some form of treatment randomization/unconfoundedness assumption. In terms of what can go wrong, for both, the presence of an additional shock or hidden treatment at the cut-off or between time 0 and time 1 can lead to a violation of the relevant continuity/parallel trend assumption.
- **Can we flip everything to get an ATC?** in principle, you could flip all the notation above to develop an identification strategy for the ATC. However, this requires a version of the parallel trends assumption that seems more questionable. We then have to assume that the trend in the treated unit reflects what would have happened if the focal unit had been treated. But often, we think of treatments as a kind of 'disruption' with possibly unknown consequences while the control is a 'baseline, do nothing' type situation. From that perspective, it seems stronger to assume that 'disrupting' the control group would have resulted in the same trend as 'disrupting' the treated group than to assume that 'leaving the treated group alone' would have resulted in the same trend as 'leaving the control group alone.'

10.1.5 Further Resources

- Cunningham (2021) has a chapter on Dif in Dif `https://mixtape.scunning.com/09-difference_in_differences`

10.2 Lagged Outcome Models: an Alternative to DiD

Another way of approaching panel data like the kind we used for DiD is to think of the pre-treatment outcome as a pre-treatment confounder and condition on it just as we would condition on other pre-treatment confounders. In this modelling approach, we replace parallel trends with an unconfoundedness assumption. This assumtion neither implies nor is implied by the DiD parallel trends assumption and ultimately, neither modelling approach is uniformly better.

10.2.1 Set-up

The basic lagged outcome model is:

$$Y_{i1}(z) = \alpha + \rho Y_{i0} + \tau z + \epsilon_{i1}(z) \tag{27}$$

where

- α is an intercept term equal to $\mathbb{E}(Y_{i1}(0)|Y_{i0} = 0)$ (notice: it is not a unit fixed effect α_i as in DiD)
- z is a treatment indicator and τ the fixed effect of treatment
- Y_{i0} is the outcome of unit i at time 0
- $\epsilon_{i1}(z)$ is an error term which is allows for heterogeneous treatment effects in the same way as the heterogeneous effect model in Module 4 Section 4.1.1

As in Module 7, the key assumption is unconfoundedness:[57]

$$Y_{i1}(1), Y_{i1}(0) \perp\!\!\!\perp G_i | Y_{i0}$$

This says that conditional on pre-treatment outcomes, there is no tendency for units with certain kinds of potential outcomes to sort into the treated and control groups. Notice that in DiD, we made no randomization-of-treatment type assumption – here we bring that in but make no parallel trends assumption. Note that this approach can be generalized to add other pre-treatment covariates X_i to the regression. The unconfoundedness assumption is then

$$Y_{i1}(1), Y_{i1}(0) \perp\!\!\!\perp Z_{i1} | Y_{i0}, X_i$$

As in Module 7, we also need the **overlap** assumption that within each Y_{i0}, X_i group, the propensity scores are not 0 or 1.

Example: consider again the tutoring example where we have pre- and post-tests give to tutored and non-tutored children. Regressing on Y_{i0} means we look at children who have the same pre-test score and then examine whether the tutored children had higher post-test scores than the non-tutored children. The intuition is that we use pre-test scores to compare children who are similar to each other and perhaps, do not differ systematically in baseline skill and prior knowledge. ϵ_i being allowed to vary between treatment and control is saying we allow the treatment and control group to have different outcome variability (as in the Neyman estimator from Module 3 and heteroskedasticity-robust standard errors in Module 4).

In this case, the unconfoundedness assumption, means that we do not, for example, have that even among students scoring very low on the pre-test, some other factor makes it that those with high potential outcomes get tutoring and those with low potential outcomes do not or vice versa, as might happen if people with higher socio-economic status have more access to tutoring and that high SES also helps them score highly on the post-test through other mechanisms. If we add socioeconomic status as an X_i in the regression above, we could also control for this confounder.

Identification under unconfoundedness and overlap is identical to that in Module 7 and 8. Once we treat Y_{i0} as a pre-treatment covariate. If the estimand is the same ATT as in DiD, we have:

$$\begin{aligned}
\mathbb{E}(Y_{i1}(1) - Y_{i1}(0)|G_i = 1) &= \mathbb{E}(Y_{i1}(1)|G_i = 1) - \mathbb{E}(\mathbb{E}(Y_{i1}(0)|Y_{i0}, G_i = 1)\ |G_i = 1) && \text{(Law of Total Expectation)} \\
&= \mathbb{E}(Y_{i1}(1)|G_i = 1) - \mathbb{E}(\mathbb{E}(Y_{i1}(0)|Y_{i0}, G_i = 0)|G_i = 1) && \text{(unconfoundedness)} \\
&= \mathbb{E}(Y_{i1}|G_i = 1) - \mathbb{E}(\mathbb{E}(Y_{i1}|Y_{i0}, G_i = 0)|G_i = 1) && \text{(consistency)}
\end{aligned}$$

which we can estimate with heteroskedastic error by fitting regressions on the $G_i = 1$ and $G_i = 0$ group separately and then calculating:

$$\frac{1}{n_1} \sum_{i=1}^{n} G_i (Y_{i1} - \hat{\mathbb{E}}(Y_{i1}|Y_{i0}, G_i = 0))$$

In terms of the linear model above, we equivalently have the following identification argument

[57] Note: this is equivalent to $Y_{i1}(1), Y_{i1}(0) \perp\!\!\!\perp Z_{i1}|Y_{i0}$. The slides have this except there is a typo where it says Z_{it} instead of Z_{i1}

$$\begin{aligned}\mathbb{E}(Y_{i1}(1) - Y_{i1}(0)|G_i = 1) &= \alpha + \rho Y_{i0} + \tau - \alpha - \rho Y_{i0} + \mathbb{E}(\epsilon_i(1)|Y_{i0}, G_i = 1) - \mathbb{E}(\epsilon_i(0)|Y_{i0}, G_i = 0) \\ &= \tau + \mathbb{E}(\epsilon_i(1)|Y_{i0}, G_i = 1) - \mathbb{E}(\epsilon_i(0)|Y_{i0}, G_i = 0) \\ &= \tau + \mathbb{E}(\epsilon_i|Y_{i0}, G_i = 1) - \mathbb{E}(\epsilon_i|Y_{i0}, G_i = 0) \\ &= \tau\end{aligned}$$

where the unconfoundedness (exogeneity assumption) implies these error terms have mean 0.

10.2.2 DiD vs Lagged Outcome

DiD and Lagged Outcome models make different assumptions and which is more plausible depends on context. Neither assumption implies the other and it is also possible for neither to hold. The key DiD assumption is

$$\mathbb{E}[Y_{i1}(0) - Y_{i0}(0)|G_i = 1] = \mathbb{E}[Y_{i1}(0) - Y_{i0}(0)|G_i = 0]$$

while the Lagged Outcome unconfoundedness implies

$$\mathbb{E}(Y_{i1}(1) - Y_{i1}(0)|Y_{i0}(0), G_i = 1) = \mathbb{E}(Y_{i1}(1) - Y_{i1}(0)|Y_{i0}(0), G_i = 0)$$

Table 6 summarizes key assumptions and ideas for the two approaches.

	DiD	**Lagged Outcome**
Assumes	parallel trends	unconfoundedness, overlap given Y_{i0}, X_i
Allows	imbalance in lagged outcome from unobserved time-invariant confounder (systematic differences) between treat and control	systematic difference in treat vs control lagged outcomes occurring because lagged outcome directly affects treatment assignment $Y_{i0} \to Z_{i1}$
Does not allow	time-variant confounders create systematic differences between treat and control that do not come from treatment	systematic differences between treat and control at time 1 that persist after conditioning on X_i, Y_{i0} because there are unobserved confounders affecting treatment assignment

Table 6: Comparison

Overall then, if your main worry is unobserved time-invariant confounders, DiD makes sense while if your main worry is the pre-treatment outcome being a confounder, lagged outcome makes sense. If both are concerns, this can be tricky.

The lagged outcome assumption could hold without DiD holding if, for example, Y_{i0} strongly causes treatment status to an extent that the treatment and control group systematically differ too much for parallel trends to be plausible but within Y_{i0}, there is still some randomization. For example, imagine those with low pre-test math scores Y_{i0} are more likely to get tutoring and differ so systematically from those with high pre-test math scores that parallel trends is implausible. However, it might still be true that once you fix pre-test score, who does and does not get tutoring is essentially random, though with higher probabilities of treatment for lower Y_{i0} values.

The DiD assumption could hold without lagged outcome holding if among those with certain Y_{i0} values, there are still systematic differences between the treated and control groups but it is still plausible that the treated and control groups would evolve in parallel absent any treatment. For example, suppose among low pre-test score Y_{i0} students, tutored students had higher SES and systematically higher potential outcomes and untutored students had lower SES and systematically lower potential outcomes. That violates the lagged outcome requirements. But if we make parallel trends assumption, we only require that if untutored students' scores increase by 20% on average in the post test, that reflects the change that would have happened to the tutored students if untutored, even if the tutored students have systematically higher scores at pre and post test.

There is a **special case** in which the lagged outcome unconfoundedness assumption implies the parallel trends assumption (though still not vice versa!): Suppose that at baseline time 0, there are no systematic differences between the two groups:

$$\mathbb{E}(Y_{i0}(0)|G_i = 1) = \mathbb{E}(Y_{i0}(0)|G_i = 0) = \mathbb{E}(Y_{i0}(0)) \tag{28}$$

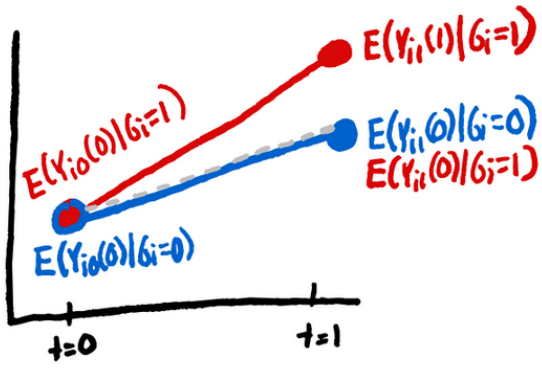


Figure 34: Special case

Then the parallel trends assumption becomes that the control group expected outcome at time 1 is the treatment group expected outcome at time 1.

$$\mathbb{E}[Y_{i1}(0)|G_i = 1] = \mathbb{E}[Y_{i1}(0)|G_i = 0] \tag{29}$$

The unconfoundedness assumption implies that for any Y_{i0}, $\mathbb{E}(Y_{i1}(0)|G_i = 1, Y_{i0}) = \mathbb{E}(Y_{i1}(0)|G_i = 0, Y_{i0})$ which, by the law of total expectation, implies Equation 29. The reverse implication (parallel trends implying lagged outcome) does not hold because parallel trends is only ever about on average quantities while the lagged outcome involves individual-level outcomes.

10.3 Fixed Effects Regressions

The data we have been discussing for DiD are a two-panel case of the more general panel data structure that is common in the social sciences. Panel data contain repeated measurements over time for the same individuals and can be a rich source of information. The data structure for panel data with T time periods and n observations can be represented as in Table 7 with a column containing a unit identifier.

Time	Unit	Treatment	Outcome
1	1	0	5
2	1	1	40
⋮	⋮	⋮	⋮
T	1	1	12
⋮	⋮	⋮	⋮
1	n	0	1
2	n	0	2
⋮	⋮	⋮	⋮
T	n	0	4

Table 7: Illustration of data structure

The key advantage of panel data is that the temporal ordering and repeat measurement of the same individual can support causal conclusions, even in the presence of *time-invariant* confounding. Regression models with fixed effects are the primary workhorse for doing causal inference on this kind of data but this raises the question of what causal assumptions are required for those regressions to actually provide estimates of causal quantities. We will briefly overview a number of different commonly used models and considerations.

Updated running example: Let's again think about tutoring and math scores but now assume that we observe students in grades 1-12. For each year, we observe whether the student was tutored or not and how the student performed on an end-of-year exam. We might also observe student socioeconomic status as a possible confounder.

10.3.1 One-Way Fixed Effect Model

The one way fixed effect model has the form

$$Y_{it} = \alpha_i + \beta X_{it} + \epsilon_{it} \quad \forall i = 1, ..., n; \quad t = 1, ..., T$$

where

- Y_{it} is the **outcome** for unit i at time t
- X_{it} is the **treatment** for unit i at time t
- α_i is a unit-specific fixed effect. This term accounts for time-invariant confounders and is only identifiable and different from the error term when we have repeated measurements for each unit.[58]
- ϵ_{it} is a time-specific error term with $\mathbb{E}(\epsilon_{it}) = 0$.[59]

The corresponding structural equation model in terms of **potential outcomes** is

$$Y_{it}(x) = \alpha_i + \beta x + \epsilon_{it}(x) \quad \forall i = 1, ..., n; \quad t = 1, ..., T$$

where

- $Y_{it}(x)$ is the potential outcome for unit i at time t when the treatment at time t is $X_{it} = x$. By consistency, the observed Y_{it} is $Y_{it} = Y_{it}(X_{it})$. Implicitly here, we are also making a **no carryover effect** (no spillover from same unit at other time periods) assumption that $Y_{it}(x_t) = Y_{it}(x_1, ..., x_t)$, which says the potential outcome is only a function of the present treatment (Imai and Kim, 2019, Assumption 2).

[58] In a regression where we only have one observation per unit, you could still write down $Y_i = \alpha_i + \beta X + \epsilon_i$ but the α_i are non-identifiable and essentially just part of the error term. If their mean is α, then we have $Y_i = \alpha + \beta X + u_i$ for $u_i = \alpha_i - \alpha + \epsilon_i$. It is only repeated measurements that make individual effects identifiable – i.e. which allow us to distinguish some unit-specific tendency from from baseline random fluctuations in individual measurements. For example, we might have a child who generally does well on math exams (even if for any one exam there are also some circumstantial factors creating fluctuations) or a patient who has baseline high cholesterol (even if the particular reading varies from doctor's appointment to doctor's appointment).

[59] As usual, if $\mathbb{E}(\epsilon_{it}) = c_i \neq 0$ we could incorporate this into the α_i term so without loss of generality we let the error term have unconditional mean 0.

- $\epsilon_{it}(x)$ is an error term which is allows for heterogeneous treatment effects in the same way as the heterogeneous effect model in Module 4 Section 4.1.1.[60]
- The estimand β is an "**average contemporaneous effect**" (Imai and Kim, 2019) of X_{it} on Y_{it}, averaged over all units and time periods. In finite sample terms, this is:

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T} Y_{it}(1) - Y_{it}(0)$$

or in super-population terms (of units, but for fixed T time periods)

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}(Y_{it}(1) - Y_{it}(0)))$$

Note, however that actually, only units with some variation in treatment over the time periods end up contributing to the estimator, so really the estimand is the above conditional on the set of units with $\sum_{t=1}^{T} X_{it} \in [1, T-1]$ (treated at least once and not at every time period) or, in a super-population sense, the units with positive probability of this. That is, if there are, for example, some kids who have no chance of ever receiving tutoring in any grade, then they are not part of our estimand.

The key assumption allowing causal identification is **strict exogeneity**:

$$\mathbb{E}(\epsilon_{it}|X_i, \alpha_i) = 0$$

where $X_i = (X_{i1}, ..., X_{iT})$.[61] This says that given a particular unit and their entire vector of treatments over time, there are no further systematic tendencies in the outcome and in a way that varies over time. For example, this assumption would be violated in the multi-year tutoring study if there were some hidden variable – such as language background – that affects both likelihood of different treatment patterns and the outcome in a way that varies over the grades. For example, perhaps students who do not have English as their first language are more likely to get tutoring in elementary school and have systematically higher/lower math scores regardless of tutoring but this confounder only matters in elementary school and later ceases to make a difference (note: this is a purely hypothetical example). If language background had a consistent effect on math scores and likelihood of receiving tutoring across all years of schooling, then it would be accounted for by a_i but otherwise, it is a time-varying confounder.

The identification argument is, letting $x_{1:T}$ be a vector of treatments at each time period and x_t be its element at time t

$$\begin{aligned}
\mathbb{E}(Y_{it}|X_{it} = x_{1:T}, a_i) &= \mathbb{E}(Y_{it}(x_{1:T})|X_{it} = x_{1:T}, a_i) \quad \text{(consistency)} \\
&= \mathbb{E}(Y_{it}(x_t)|X_{it} = x_{1:T}, a_i) \quad \text{(no carryover)} \\
&= \alpha_i + \beta x_t + \mathbb{E}(\epsilon_{it}(x)|X_{it} = x_{1:T}, a_i) \\
&= \alpha_i + \beta x_t \quad \text{(exogeneity)}
\end{aligned}$$

In the case of a binary treatment, letting $X = (X_1, ..., X_n)$ where each $X_i \in \mathbb{R}^T$ and $\alpha = (\alpha_1, ..., \alpha_n)$ and assuming independence across observations, it follows that, for $x = 0, 1$:

$$\mathbb{E}(\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} I(X_{it} = x)Y_{it}|X, \alpha) = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} a_i + \beta x$$

and hence the following is an unbiased estimator for β under the exogeneity assumption

$$\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} X_{it}Y_{it} - (1 - X_{it})Y_{it}$$

this should look very familiar. It is the same difference in means type result we had from Neyman in Module 3 and translated to regression in Module 4.

[60] Quick recap: if we had just ϵ_{it}, we'd be posing constant individual-level effects. We could instead write down this model with a $\beta_{it} - \beta$ term where $\mathbb{E}(\beta_{it}) = \beta$. The error term then becomes $\tilde{\epsilon}_{it}(x) = (\beta_{it} - \beta)x + \epsilon_{it}$, incorporating the individual heterogeneity. Assuming a constant vs heterogeneous effect model does not make a difference for identification because though individual-level effects may vary across units and over time, we can ultimately only identify an average over units and time periods. We *could* also identify an average effect at just one time point (as we did in previous modules) if we could control for appropriate confounders but here we use the repeated measurements to do that.

[61] Note there is a slight notational ambiguity here. α_i looks like a parameter but is being conditioned on here as if it is a random variable. We could define n random variables I_{ij} that are indicators for whether Y_i comes from unit j. Technically, our regression has these n parameters in it and we would then say $\mathbb{E}(\epsilon_{it}|X_i, I_i)$ where $I_i = (I_{i1}, ..., I_{in})$. But because this gets very cumbersome, we essentially use α_i as that indicator here. So the statement is "conditional on our observations coming from unit i"

The **non-parametric generalization** of the above set-up is represented in the DAG in Figure 35. As discussed in Section 7.6.2, the DAG can be thought of as a non-parametric structural equation model. The linear model above is then a special case that assumes linearity in treatment (if it is continuous) and additive error. More generally, we have

$$Y_{it} = g_1(X_{it}, U_i, \epsilon_{it})$$

$$X_{it} = g_2(X_{i1}, ..., X_{i,t-1}, U_i, \eta_{it})$$

where ϵ_{it}, η_{it} are independent noise terms.

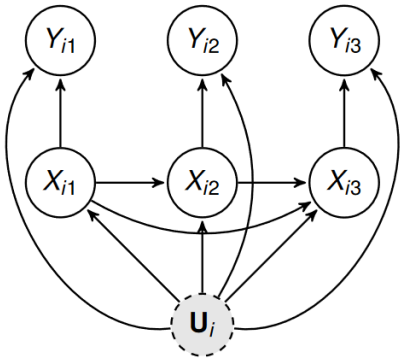


Figure 35: DAG for generalization of one-way fixed effects model from Slide 11

Here U_i represents the time-invariant confounder(s). The DAG and non-parametric model connect to the previous linear model as follows:

- In the linear model, we had a linear unit-specific fixed effect a_i that accounted for time-invariant confounders that might make a certain unit consistently have a higher/lower outcome for reasons apart from treatment. Implicitly we had $a_i = h(U_i)$ for some unknown function h that we did not need to know. Here, we generalize this to just posing that there are some time-invariant confounders U_i.
- In the linear model, we did not allow for any time-varying confounders. This is also reflected in the DAG by the absence of any U_{it} pointing to X_{it} and Y_{it} alone.
- We still assume no carryover effect $Y_{it}(x_{1:T}) = Y_{it}(x_t)$
- Neither the DAG nor the linear model precludes X_{it} at time t being affected by any previous treatment statuses. For example, you could have a randomized experiment in which treatment at time t is randomized at time t but subject to some constraints based on $X_{i1}, ..., X_{it-1}$ such as that no unit can be treated more than thrice or no unit can have treatment more than twice in a row etc.

Overall, the DAG encodes the following assumptions:

1. **No unobserved time-varying confounder exists**
2. **Past treatments do not directly affect the current outcome.** They may indirectly via their effect on future treatments. If we did have a $X_{i1} \to Y_{i2}$ arrow in the DAG above, then X_{i1} would be playing the role of a confounder relative to $X_{i2} \to Y_{i2}$.
   Example: violated if, even among students tutored in grade 4, students tutored in grade 3 did better on the math exam in grade 4
3. **Past outcomes do not affect current outcomes**

   Example: violated if scoring badly on an exam in grade 3 hurts students confidence so that it directly causes a worse outcome on an exam in grade 4, even while holding tutoring status in each year constant.

4. **Past outcomes do not affect the current treatment**
   Example: violated if exam grade in grade 3 affects whether tutored in grade 4 (very realistic in this scenario!)

See Imai and Kim (2019) for more. In particular, the paper formalizes these assumptions in potential outcomes notation and gives more information about identification. Slides 12-15 consider different cases where some of the above assumptions are violated. In some cases, identification is still possible. Again, see Imai and Kim (2019) for a more thorough exposition.

10.4 Synthetic Control

We can leverage synthetic control methods when we have panel data with a pre-treatment time period. In the simplest case, we have a single treated unit that receives treatment after some number of time periods while the other control units never receive treatment. We then use the data from the pre-treatment time period to learn how the control outcomes correlate with the treated units' outcomes and, much like in DiD, use this relationship to infer what the treated unit's outcome would have been if not treated. The intuition is that, especially if we observe a correlation between control and treated unit outcomes over a long pre-treatment period, the correlation is likely to reflect underlying commonality on unobserved pre-treatment factors that will still apply in the post-treatment period. That said, as in DiD, the control units may be systematically different from the treated unit – the key assumption is that there is some correspondence between their trends. As in Module 8 (matching and weighting), we want to match with control units which are in some way similar so that parallel trends type assumptions are more plausible. We then weight the control units so that they look like the treated unit's outcomes in the pre-treatment period and apply those learned weights to the post-treatment period. These ideas can be extended to cases where multiple units are treated at possibly different times but we focus first on the single treated unit case.

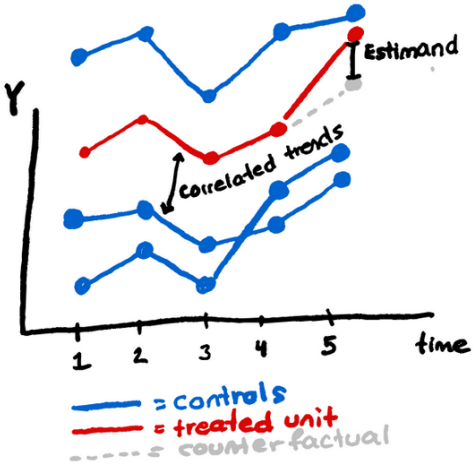


Figure 36: Illustration of basic synthetic control intuition.

10.4.1 Examples

Example 0: going back to the idea of studying the effect of tutoring on math scores in a case where we have observations over time, suppose we focus on a particular student who only received tutoring starting in 9^{th} grade. If we have a group of students who received no tutoring up to and including 9^{th} grade, then we might use their math scores in 9^{th} grade to try to infer the focal student's score under no-tutoring. In particular, we'd look to up-weight students' whose scores closely tracked with the focal student's scores in the past. This approach assumes that the correlation between students stays constant. That correlation might be threatened, for example, by the fact that in the US, 9^{th} grade is often the transition point between middle and high school. Therefore, other things may have happened between 8^{th} and 9^{th} grade that both caused the focal student to get tutoring while others did not and are correlated with the potential outcomes for the 9^{th} grade math test.

Example 1: The original example of a synthetic control estimator is Abadie and Gardeazabal (2003). Briefly, this study examines the economic impact of conflict via the case of terrorism in the Basque region of Spain by the ETA separatist group starting in the 1960s. The key idea is to look at economic trends in the the Basque Country and other regions of Spain prior to the start of the conflict (see plot Slide Slide 3). Those regions might not have the same exact economic numbers, but the idea is that their general trends have some relationship which we can learn and use to construct a 'synthetic control' for the Basque Country. This analysis relies on the assumption that absent any conflict, the relationship between the economies in other regions and in the Basque Country would stay the same. It also assumes there is no spillover in which the conflict in Basque also affects the economies of the other regions. See also Chapter 10 of Cunningham (2021) for more discussion of this example.

Example 2: Abadie et al. (2010) consider the case of Proposition 99, which increased the cigarette consumption tax and funded other anti-smoking initiatives in California in the year 1988. In some other, control states, there was no such program. The paper examines per-capita cigarette sales. Intuitively, if these decline in California but not in other states, we might suspect an effect of the proposition. But in fact, cigarette sales were already decreasing, even prior to the Proposition. If we use other states without a major anti-cigarette program passed around that time to form a synthetic control, then we can account for this possibility. See also Chapter 10 of Cunningham (2021) for more discussion of this example.

10.4.2 Basic Synthetic Control Set-up, Estimand, Identification, Estimation

In the simplest case, the synthetic control set-up is as follows:

- N units
- Each unit observed at T times. Note that we do not assume that times $1, ..., T$ are evenly spaced per se. Instead, time is ordinal categorical here. That is, we have some times $t_1 <, ..., < t_T$ indexed by $t = 1, ..., T$. Note also that these times are treated as **fixed**. They are not random times.
- At times $t = 1, ..., T-1$ no units are treated
- At time T **one unit** has been treated and the other units (control units) have not. For notational simplicity, let the treated unit be indexed by $i = N$.
- The potential outcome of unit i at time t with treatment x is $Y_{it}(x)$. Assuming consistency, the set-up implies that $Y_{it} = Y_{it}(0)$ for $i = 1, ..., N;\ t = 1, .., T-1$ while $Y_{NT} = Y_{NT}(1)$ and $Y_{iT} = Y_{iT}(0)$ for $i \neq N$.

The **estimand** is the causal effect on the treated unit N at time T:

$$\tau_N = Y_{NT}(1) - Y_{NT}(0) = Y_{NT} - Y_{NT}(0)$$

Note that this is not an average effect. It is an individual effect for which $Y_{NT}(1)$ is observed but $Y_{NT}(0)$ must be estimated. If $T = 2$ and we have only two units, then this immediately reduces to the DiD set-up.

To **identify** this estimand, we first make a **Convex Hull Assumption**, which says that there exist some true, non-negative weights $w_1, ..., w_{N-1}$ which sum to 1 such that the unknown $Y_{NT}(0)$ for the treated unit is equal to a weighted average of the other control units' outcomes: $\sum_{i=1}^{N-1} w_i Y_{iT}(0) = Y_{NT}(0)$.[62] The key idea of synthetic controls is that we can learn these weights from the pre-treatment outcomes $\{Y_{it}, t = 1, ..., T-1, i = 1, ..., n\}$, which tell us how the focal treated unit correlates with the pre-treatment units in the absence of any treatment. Intuitively, even if unit N tends to have systematically different outcomes from the controls as in Figure 36, as long as its correlations can be extrapolated to the post-treatment potential outcome under no treatment, we can use a relationship learned on the pre-treatment period to infer the missing counterfactual. Formally, we minimize the sum of squared differences between the treated unit pre-treatment observation Y_{Nt} and the weighted average control unit observation, summed over all pre-treatment time periods:

$$\hat{w} = \arg\min_{w \in A} \sum_{t=1}^{T-1} \left(Y_{Nt} - \sum_{i=1}^{N-1} w_i Y_{it} \right)^2$$

This is minimized subject to the constraint that w are in the set A of weights which are non-negative and sum to 1:

- $\sum_{i=1}^{N-1} w_i = 1$
- $w_i \in [0, 1]$

These constraints help prevent extrapolation by making sure that the imputed $\sum_{i=1}^{N-1} \hat{w}_i Y_{iT}$ is in the convex hull formed by the $Y_{1T}, ..., Y_{N-1,T}$. That is, the constraints ensure that $\min_{i \in \{1,...,N-1\}} Y_{iT}(0) \leq \sum_{i=1}^{N-1} w_i Y_{iT}(0) \leq \max_{i \in \{1,...,N-1\}} Y_{iT}(0)$. Overall then, to **estimate** $Y_{NT}(0)$, we create a synthetic control by applying the estimated weights to the control unit outcomes at time T:

$$\widehat{Y_{NT}(0)} = \sum_{i=1}^{N-1} \hat{w}_i Y_{iT} = \sum_{i=1}^{N-1} \hat{w}_i Y_{iT}(0)$$

The final **synthetic control estimator** is

$$\hat{\tau}_N = Y_{NT} - \widehat{Y_{NT}(0)} = Y_{NT} - \sum_{i=1}^{N-1} \hat{w}_i Y_{it}$$

In addition to the Convex Hull assumption, the above identification makes the following two assumptions, which we state below a bit informally (See Section 10.4.9 for attempts to formalize the justification for synthetic control).

- **Ability to learn accurate weights from controls in pre-treatment period**: even if the convex hull assumption holds so that true weights exist for time T, the weights $\hat{w}$ we get from the optimization problem above may not be equal to those true weights. This can be due to the control units exhibiting some noise in their trends or from a failure of extrapolation where the treated-control relationship observed in the pre-treatment period does not generalize. See Section 10.4.6 for more discussion of bias and variance. Overall, we are making an assumption somewhat like that

[62] This assumption is violated if the treated unit's outcomes are systematically higher (or lower) than those of all the control units. In this case, there might still be correlations that we could, perhaps reasonably, extrapolate from. As discussed further in Section 10.4.5, it is sometimes still be possible to use a synthetic controls type approach in this situation.

of parallel trends in DiD. Substantively, the things to worry about are (1) major structural changes or non-treatment shocks to the treated and/or control between the pre- vs post-treatment window and (2) the control units not having any real relationship to the treated units, e.g., if the correlation observed in the pre-treatment period is really entirely spurious.

- **No spillover**: we assume that the treatment of unit N does not affect the outcomes of units $i = 1, ..., N-1$ so that they truly reflect what would happen without treatment. That is, letting $Z_1, ..., Z_N$ be the treatment assignments for the N units, we assume $Y_{iT}(z_i = 0, z_N = 1) = Y_{iT}(z_i = 0, z_N = 0) = Y_{iT}(z_i = 0)$

Note that we could also apply these weights to calculate an estimate of the effect at times $T+1, T+2, ...$ However the synthetic control assumptions may become less tenable the larger the gap in time between the period we used to learn the weights and the $Y_{Nt}(0)$ we want to estimate.

10.4.3 Placebo Tests

Given a synthetic control estimate, we might also ask how to evaluate whether the true causal effect is different from 0 or more generally, to quantify the uncertainty in the estimate. One approach is to use the **placebo test**, which is also a way to examine whether the method is behaving as expected. The steps are:

1. Pick a focal control unit j
2. Run synthetic controls procedure using $Y_{i,t}$ for $i \neq j$ and $t = 1, ..., T-1$ as the control unit observations (note: we can include the observations for unit N since these are also under control for $t < T$)
3. Calculate $\hat{\tau}_j = Y_{j,T} - \sum_{i \neq j} w_i Y_{i,T}$

since unit j is not treated, $\hat{\tau}_j$ is really an estimate of $Y_{j,T}(0) - Y_{j,T}(0) = 0$ and hence if the synthetic control is working well, the estimated difference from step 3 above should be small. From a testing perspective, if we are willing to assume that the assumptions of synthetic control do hold, repeating the above procedure for different focal control units gives us an indication of how large the estimated treatment effect for our original focal unit $Y_{NT}(1) - \widehat{Y_{NT}(0)}$ could plausibly be just by chance even if $Y_{NT}(1) = Y_{NT}(0)$ (no true effect).

Formally, let $\hat{\tau}_j$ be the placebo effects from the procedure above for control units $j = 1, ..., N-1$. Assume that $\hat{\tau}_j$ are exchangeable, which essentially means that although they are not independent because they are calculated from common data, they are identically distributed and indistinguishable in terms of their joint distribution.[63] Under the assumptions of synthetic control, these quantities reflect some baseline level of random noise. The null hypotheses is $H_0 : Y_{NT}(1) = Y_{NT}(0)$, which would make $\hat{\tau}_N$ just another draw from the distribution that generated $\hat{\tau}_j$. We can then calculate a p-value as:

$$p = \frac{1}{(N-1)+1}\left(1 + \sum_{j=1}^{N-1} I(|\hat{\tau}_j| \geq |\hat{\tau}_N|)\right)$$

This is a form of **permutation test** as in Module 2 in the sense that we are considering all the different 'permutations,' each resulting in a different unit being treated as the focal unit. Depending on the context, it may also make sense to make this p-value one-sided. See also Cunningham (2021) Chapter 10 for a related approach which relies on a mean squared error calculation rather than simply counting how many are more extreme. As in Module 2, we can also invert this test by specifying different null values $Y_{NT}(1) - Y_{NT}(0) = \tau_0$, adjusting our synthetic control estimator to reflect this

$$\hat{\tau}_N(\tau_0) = (Y_{NT} - \tau_0) - \widehat{Y_{NT}(0)}$$

running the test, and collecting a set of τ_0 for which we do not reject.

Figure 37 shows what this looks like in the California Proposition 99 example from Abadie et al. (2010). We see that, especially right after 1988, where the synthetic control assumptions are most plausible, California is quite an outlier in its estimated negative effect. As we move farther away from the 1988 cut-off, there is more variability and synthetic control assumptions are less plausible, so it is harder to draw confident conclusions here. Still, California remains an outlier.

[63] Formally, if we have n random variables $X_1, ..., X_n$ with joint distribution p, these are (finite) exchangeable if $p(X_1, ..., X_n) = p(X_{\pi(1)}, ..., X_{\pi(n)})$ for any permutation π. https://en.wikipedia.org/wiki/Exchangeable_random_variables

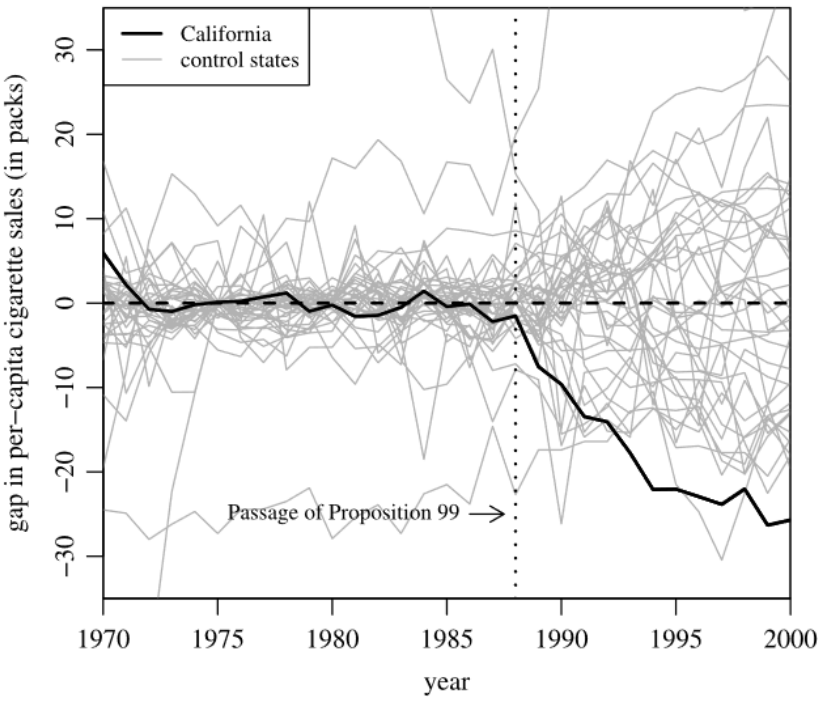


Figure 37: Figure from Slide 7

10.4.4 Standard Errors for Synthetic Control

Notice that even for the synthetic control method run on placebo states in Figure 37, the synthetic control estimate is not always exactly 0 (its true value). As usual in statistics, there are two sources of error we worry about for synthetic control methods: bias and variance. We discuss bias in Section 10.4.8 below and focus on variance first. First note that the variance of the synthetic control estimator $\hat{\tau}_N$ is generally considered to depend only on the estimation of $Y_{NT}(0)$ while $Y_{NT} = Y_{NT}(1)$ is treated as fixed.[64] To think about variance in our estimate of $Y_{NT}(0)$, we need to answer the question, "where does the statistical randomness come from?" In synthetic controls, this isn't immediately straightforward to think about.

If we are willing to pose one of the models discussed at the end of this section as ways of justifying synthetic controls more formally (Section 10.4.9), then this may also come with assumptions about standard errors and ways of estimating them. Without these models, one approach is to take a **design-based** approach. From this perspective, we think about the fact that **Variance** can arise from the fact that the control unit trends over times $t = 1, ..., T-1$ may reflect a noisy trend with some variations arising due to systematic factors they (hopefully) have in common with the treated unit and other variations arising due to other particular circumstances for that unit (Figure 38). In some settings, it might also make sense to think of the control units as some form of random sample, again with some baseline level of noise in their trends, even if those trends are correlated with that of the treated unit. The placebo test procedure described in the previous section suggests one way to estimate this baseline variance.

As described earlier, the variability in placebo estimates $\hat{\tau}_j$ around true value 0 reflects baseline noise arising from the individual fluctuations of particular control units. Hence, one variance estimator is:

$$\hat{\mathbb{V}}(\widehat{Y}) = \frac{1}{N-1} \sum_{j=1}^{N-1} \hat{\tau}_j^2$$

See Doudchenko and Imbens (2016) for more on this *design-based* approach. Other variance estimation approaches exist based on conformal inference and modelling exist. See Victor Chernozhukov and Zhu (2021); Pang et al. (2022).

[64] Strictly speaking then, we are only estimating an effect on a single unit at a single time and if we wanted to argue that effect would apply to other similar units or to the same unit at a different time, that requires an external validity argument and might not be very credible.

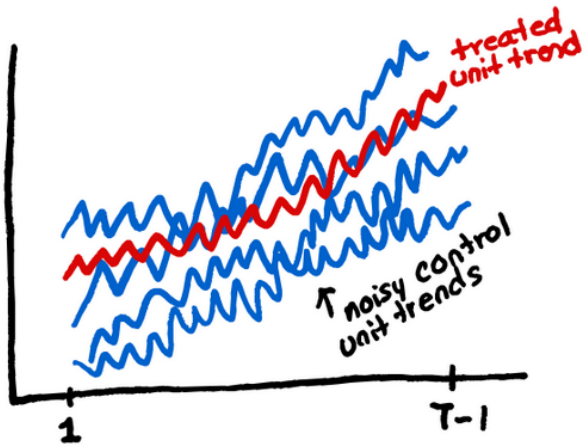


Figure 38: Illustration of noisy control unit trends

10.4.5 Synthetic Control and Regression

It turns out that synthetic control can also be viewed as a regression of the treated unit outcomes on the $N-1$ control unit outcomes with each control unit a covariate subject to some constraints. That is, in the usual regression terminology, we have:

- **Outcome vector** $Y_N = (Y_{N1}, ..., Y_{N,T-1})$, length $T-1$ vector
- **Design matrix**: $T-1 \times N-1$ matrix where each 'covariate' is a control unit and we have $T-1$ observations for that covariate.

$$\begin{bmatrix} Y_{11} & ... & Y_{N-1,1} \\ \vdots & \ddots & \vdots \\ Y_{1,T-1} & ... & Y_{N-1,T-1} \end{bmatrix}$$

- **Coefficients**: $(\alpha, w_1, ..., W_{N-1})$ parameter vector containing a weight for each control unit (covariate). α represents the intercept variable
- **Errors**: $(\epsilon_1, ..., \epsilon_{T-1})$ error terms representing unaccounted for variation in Y_{Nt} at each time period

Figure 39 illustrates this data structure, as well as the missing value we want to predict (it isn't missing under treatment, but it's missing under control).

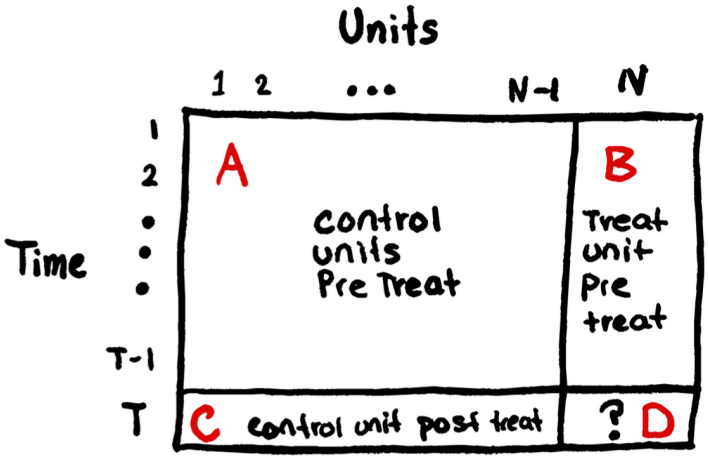


Figure 39: Illustration of data structure that could be used to fit a regression synthetic control type data

Standard OLS regression then minimizes

$$\arg\min_{\alpha,w} \sum_{t=1}^{T} \left(Y_{Nt} - \alpha - \sum_{i=1}^{N-1} w_i Y_{it} \right)^2$$

This is identical to the synthetic controls minimization problem if we add the constraints that:

1. $\alpha = 0$
2. $w_i \in [0, 1]$
3. $\sum_{i=1}^{N-1} w_i = 1$

This equivalence highlights the role that each of these constraints plays and what would happen if we loosened them:

1. If we allow $\alpha \neq 0$, then we allow there to be some time-invariant systematic difference between the treated and control units. This could be plausible in some scenarios such as in Figure 40. In that case, the convex hull assumption becomes that if you adjust every Y_{Nt} observation down by α, *that* value is in the convex hull of the controls. Note also that difference in differences is equivalent to the case with $N = 2$ (so $w_1 = 1$) and $\alpha \neq 0$ (time-invariant confounder).

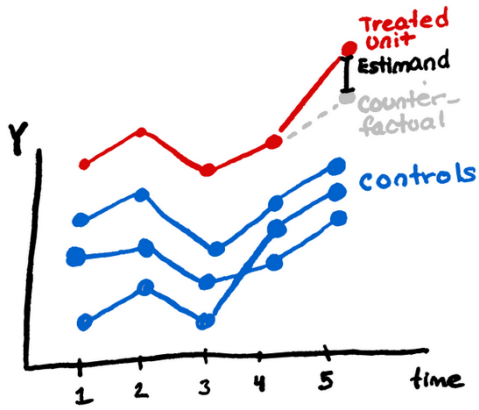


Figure 40: Illustration of $\alpha \neq 0$ type situation

2. $\sum_{i=1}^{N-1} w_i = 1$ and $w_i \in [0, 1]$ prevent extrapolation as discussed earlier. They also play a regularization role. This is a regression scenario where it is quite plausible that the number of covariates $N - 1$ might be larger than the number of replicates $T - 1$, in which case the regression cannot be solved without regularization. This points to the possibility of using other regularization constraints such as LASSO or elastic net.

Given some regression that produces a vector of weights w, the synthetic control estimate of $Y_{Nt}(0)$ in each year $t = 1, ..., T$ can be calculated as

$$\widehat{Y_i(0)} = \begin{bmatrix} Y_{11} & \dots & Y_{N-1,1} \\ \vdots & \ddots & \vdots \\ Y_{1,T-1} & \dots & Y_{N-1,T-1} \\ Y_{1,T} & \dots & Y_{N-1,T} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_{N-1} \end{bmatrix}$$

which results in a length T vector.

10.4.6 Overfit and Regularization

What if we just fit the OLS regression above without constraining the weights? In some cases, this will fail because of a dimensionality problem! In particular, if there are more control units than time periods ($N-1 > T-1$), then the least squares regression does not have a unique solution! The classic solution to this is **regularization,** which provides an alternative to constraining the weights. Regularization can also be important even if $N \leq T$ as a way to prevent overfitting to particular fluctuations in the control units' trends that are unrelated to any underlying attributes they have in common with the treated unit of interest. The danger of overfitting is that we may select weights based on patterns which do not generalize to time T. If we consider the noise in the trend as random (something that might randomly vary across repeated sampling, time periods, or measurements), such overfitting will not create systematic bias but will increase the variance in our synthetic control estimator. We see some suggestion of this in Figure 37. The tendency for the noise to grow after the treatment period cut-off in that plot (despite the true effect being 0) likely reflects synthetic control's tendency to overfit the noise in the pre-treatment data.

There are many general methods for regularization, including LASSO, Ridge Regression, and Elastic net (you will explore two of these in the Module 10 problem set). Another approach proposed by Abadie and L'Hour (2021) in the case of synthetic controls case is to minimize

$$\arg\min_{w} \underbrace{\sum_{t=1}^{T-1}\left(Y_{Nt} - \sum_{i=1}^{N-1} w_i Y_{it}\right)^2}_{\text{component wise fit}} + \lambda \underbrace{\sum_{i=1}^{N-1} w_i \sum_{t=1}^{T-1}\left(Y_{Nt} - Y_{it}\right)^2}_{\text{aggregate fit}}.$$

The first term above is our usual synthetic control optimization term and the second is a **penalty** which incentivizes larger weights for control units with pre-treatment outcomes that are actually close to that of the treated unit (small $(Y_{Nt} - Y_{it})^2$). The standard synthetic control does not actually require this. The λ term is a **tuning parameter** that controls how strong the penalty is. If $\lambda = 0$, we have the basic synthetic control method. If $\lambda \to \infty$, then we have essentially a nearest neighbor matching approach where we assign all the weight to the control unit that is overall closest to the treated unit in its outcomes, without any chance to leverage multiple controls and their trends. One approach to choosing λ is to use a train/test split, hold-out approach on the pre-treatment data, using the 'test' pre-treatment period to evaluate the different λ choices

As an example of what this regularization is doing, suppose the states of New York (NY) and Texas (TX) have very different outcomes and trends from California but NY tends to be below CA while TX tends to be above CA in such a way that their weighted average matches CA well in the pre-treatment period. The standard synthetic control method could still end up assigning CA and TX relatively large weights while the regularization above creates a pressure against this. One intuition for why this is a good idea is that NY and TX being far in the outcomes from CA suggests that they have some systematic differences from CA that could make their particular fluctuations in outcomes less likely to generalize – we are more worried about overfitting them than overfitting a state with outcomes that closely track with the actual pre-treatment outcomes of CA.

Is there a way around having to do regularization in the case of $T < N$? Interestingly Shen et al. (2022) show that an alternative regression in the $T < N$ case where we regress the observations at time T on the observations at time $1, ..., T-1$ (each a covariate, with $N-1$ replicates), obtain a coefficient (weight) for each time period, and apply these to the $\{Y_{Nt}\}_{t=1,...,T-1}$ to predict $Y_{NT}(0)$ not only works but is in a sense equivalent to the regression described earlier! If we look at Figure 39 again, we can think of these as a 'horizontal regression' (of panel B on panel A) and a 'vertical regression' (of panel C on panel A). It is pretty startling (we think) that these turn out to be the same!

10.4.7 Adding Covariates

Synthetic controls can be generalized in various ways. One is to include pre-treatment covariates X_{it} measured at each time t and learn weights while trying to balance not only the Y_{it}'s from the pre-treatment period but these covariates. As in Module 8, we really care only about balancing covariates with some relationship to the outcome. Intuitively, when creating our synthetic control, if our focal unit N is, say, a 52-year old woman from Medford Massachusetts, we might want to up-weight the control units which are also women, also in their 50s, and/or live nearby so that their outcomes contribute more to our synthetic one for the person of interest.

Formally, let $Y_i = (Y_{i1}, ..., Y_{i,T-1})^T \in \mathbb{R}^{T-1}$ be the vector of all lagged (aka pre-treatment) outcomes. Suppose at each time point, we have a p-dimensional covariate vector X_{it} and let $X_i = (X_{i1}, ..., X_{i,T-1})^T$ be length $p(T-1)$ which stacks these into a single column vector of all lagged covariates.[65] Let $Z_i = (Y_i^T, X_i^T)^T$ be the full stacked vector of pre-treatment outcomes and covariates at each time for unit i. The synthetic control optimization problem then becomes to learn weights which minimize the Mahalanobis distance between Z_N and a weighted average of the Z_i's for $i \neq N$:

$$\hat{w} = \arg\min_{w} \left(Z_N - \sum_{i=1}^{N-1} w_i Z_i\right)^T \hat{\Sigma}^{-1} \left(Z_N - \sum_{i=1}^{N-1} w_i Z_i\right) \tag{30}$$

[65] Note: if X_{iT} is measured before treatment, this can be included as well.

where $\hat{\Sigma}^{-1}$ is the inverse of an estimated of variance-covariance matrix $\hat{\Sigma}$ of Z_i. Incorporating $\hat{\Sigma}$ means the distance metric accounts for linear redundancies among the covariates. The optimization problem is again subject to the constraints that $\sum_{i=1}^{N-1} w_i = 1$ and $w_i \in [0, 1]$, though again this could be replaced with other regularization approaches.

10.4.8 Bias Corrections

When we use the weights from the pre-treatment period to estimate a treatment effect, we are assuming that they capture an underlying signal which still applies in the post-treatment time period. Bias in estimating $Y_{NT}(0)$ can arise for a few reasons. One source of bias could be if the true $Y_{NT}(0)$ does not lie in the convex hull of the $Y_{iT}(0)$. As discussed in Section 10.4.5, a time-invariant systematic difference α between the treated unit and the control units that violates the convex hull assumption is not hard to adjust for. However in general, biasing dynamics may be more complex. Bias can occur when the relationship between $Y_{Nt}(0)$ and $Y_{it}(0)$ in the pre-treatment period does not generalize to the post-treatment period. This could occur because of some shock (hard to adjust for) but also because of systematic differences between the treated unit and control unit other than treatment, e.g., because the nature of the control units' covariates makes it impossible to weigh them all perfectly in equation 30. If bias is because of covariate imbalance, then it may be possible to do a bias correction very similar to the bias correction for matching discussed in section 8.1.7.

Specifically, suppose we have some model $Y_{it}(0) = \mu_{it} + \epsilon_i$. This μ_{it} could come from the factor-analytic model or auto-regressive model discussed in the next section or even from a OLS regression of Y_{it} on X_{it}.[66]. We can then consider the following two quantities:

- $\hat{\mu}_{NT}$: the predicted value of $Y_{NT}(0)$ under this model, based directly on the covariates (time-invariant or time-variant, depending on the model) of treated unit N
- $\sum_{i=1}^{N-1} w_i \hat{\mu}_{iT}$: where w_i are the synthetic control weights and $\hat{\mu}_{iT}$ are the predicted value of $Y_{it}(0)$ for the control units given their own covariates.

Notice that if the treated and control units all have exactly the same covariates, then these two values are exactly equal. More generally, if the covariates of the control units generally look a lot like that of the treated unit, we might expect the two fitted values to not be so different. If, on the other hand, there is a remaining meaningful imbalance, then the two fitted values above may be different and their difference can capture the bias that arises from this covariate imbalance. The **augmented synthetic control estimator** incorporates that difference as a bias correction:

$$\widehat{Y_{NT}(0)} = \sum_{i=1}^{N-1} w_i Y_{iT} + \left(\hat{\mu}_{NT} - \sum_{i=1}^{N-1} w_i \hat{\mu}_{iT} \right) = \hat{\mu}_{NT} + \sum_{i=1}^{N-1} w_i (Y_{iT} - \hat{\mu}_{iT})$$

The first way of writing this estimator is very similar to the bias-corrected estimators we saw in Module 8 Section 8.1.7. We have our original synthetic control estimator and then we add on the difference in fitted values. The second way of writing the estimator on the right shows that we can also think of this as taking our fitted value $\hat{\mu}_{NT}$ (itself also directly an estimate of $Y_{NT}(0)$) and adding in a term that accounts for how actual Y_{iT} values tend to differ from their $\hat{\mu}_{iT}$ and predicts what that might look like for unit N. See Eli Ben-Michael and Rothstein (2021) for more details on the these methods. Note that these bias correction approaches come at the cost of further modelling assumptions. Results will depend on the choice of model and if the model is very wrong, the bias correction may not help.

Example: To make the above ideas a bit more concrete, consider the following hypothetical scenario: Suppose we have a collection of hospitals observed over time. The outcome of interest is their covid mortality rate, the treatment is some change in hospital procedure, and we observe various pre-treatment covariates, including the average patient age at each hospital. Suppose we focus on a particular city hospital of interest that received treatment at time T and use synthetic controls to impute $Y_{NT}(0)$. We might have selected a group of controls that are already fairly similar to the hospital (e.g., are also in cities in the same region) or our weights might already up-weight those control hospitals. But now imagine that we notice that the average patient age at our focal hospital is 10 years older than at the other hospitals. Since higher age is associated with greater covid mortality, we might be worried that there will be some bias from using control hospitals that tend to have younger patients. The bias correction idea would be to fit a model of mortality given age and other covariates (using the control units only), predict mortality for the focal hospital and the control hospitals, and use that difference as a correction. It might be that actually, $\sum_{i=1}^{N-1} w_i \hat{\mu}_{iT}$ is pretty good (close to $\hat{\mu}_{NT}$) and then we do not do much correction. But if, for example, $\sum_{i=1}^{N-1} w_i \hat{\mu}_{iT}$ ends up being a bit too low compared to what we would predict $\hat{\mu}_{NT}$ to be based on its average patient age, we end up shifting our estimate of $\widehat{Y}_{NT}(0)$ up a little.

[66] Stat 286 students will explore one a option in the Module 10 problem set Question 2

10.4.9 Model-Based Justifications for Synthetic Control

The description of synthetic controls in the previous sections does not define any parametric model. This can make it hard to formally write down the assumptions or justify the method. Although these methods were initially proposed without a model-based justification, some have sought to make them more rigorous by finding such justifications. Synthetic control methods were originally formalized by Abadie et al. (2010), who gave two possible motivating models. Keep in mind these are *post-hoc* justifications of the method. Given you believe either of these models, you might adopt a synthetic control strategy but other models could also justify the method.

1. **Factor-analytic model:** time-invariant covariates, time-varying coefficients

$$Y_{it}(0) = \gamma_t + \delta_t^T X_i + \xi_t^T U_i + \epsilon_{it}$$

where

- $Y_{it}(0)$ is the outcome under control at time t
- γ_t is a time fixed effect for time t (e.g., outcomes tending to be higher at time 4 in Figure 36)
- δ_t^T are time-varying effects of time-invariant covariates X_i. Note that X_i are not indexed by time and hence are time-invariant covariates and in particular, do not include the lagged outcomes $Y_{i,t-1}$. The effect of having covariate X_i is allowed to vary over time however.
- ξ_t^T are time-varying effects of unobserved covariates U_i
- ϵ_{it} is remaining time-unit level variation

Notice that compared to the fixed effects regressions considered previously, this set-up allows even more dynamism over time because we can have time-varying effects on outcome for both observed and unobserved confounders. The key assumption is

$$\exists w_1, ..., w_{N-1} : \quad \sum_{i=1}^{N-1} w_i X_i = X_N \quad \text{and} \quad \sum_{i=1}^{N-1} w_i U_i = U_N$$

That is, this model highlights the assumption that there exists weights which make the control group like the treatment group *in terms of observed AND unobserved variables* so that balancing on observed variables is enough to balance on unobserved ones. Built into this model is the assumption that the covariates X_i and U_i account for variation in $Y_{it}(0)$. If they did not, we would not care about balancing them.

2. **Auto-regressive model:** unconfoundedness with lagged outcomes

$$Y_{it}(0) = \rho_t Y_{i,t-1}(0) + \delta_t^T X_{it} + \epsilon_{it}$$

$$X_{it} = \lambda_{t-1} Y_{i,t-1}(0) + \Delta_{t-1} X_{i,t-1} + \nu_{it}$$

Line 1 of this model poses that the current outcome depends on past outcome and the current covariates, with effects that vary with t and **Line 2** poses that the current covariates depend on past outcome and past covariates. This is a lagged-1 model because there is only dependence on the previous time period. X_{it} are now time-varying covariates with time-specific coefficients. The key assumption made by this model is the *absence* of a U_i. There are no unobserved time-invariant confounders.

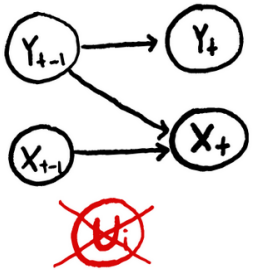


Figure 41: Segment of a DAG for the autoregressive model

Overall, the factor-analytic model is a more **static** model which only factors in time-invariant covariates, though it allows their coefficients to vary. The nice thing is that it makes explicit what we need to assume about unobserved time-invariant covariates in this set-up. The auto-regressive is more dynamic because it allows time-varying covariates, but it does not incorporate unobserved confounders. As shown in Abadie et al. (2010), synthetic control methods can be justified under either model. This is interesting because it means you do not necessarily have to commit to which one is true...but that said, if both are true (have unobserved time-invariant confounders and dynamics from auto-regressive model) then synthetic controls will not identify the correct quantity. This reflects a general **trade-off** between models that account for unobserved unit-specific time-invariant confounders and models that allow for dynamic scenarios with previous outcomes affecting current treatment – you generally cannot have both and still have an identifiable model (Video lecture; around 12:50).

10.5 Staggered Adoption

So far, we have only discussed DiD and synthetic control methods when we are interested in the treatment effect on a single treated group or even single focal treated unit. This may be realistic when units are states as in the Proposition 99 example, but there are many contexts where we might want to know about the effects for multiple units or for multiple groups of units treated at different times. This is the staggered adoption scenario, and it is quite common in practice. For example, Wood et al. (2020) study the impact of a procedural justice police training program on policing outcomes. In the study, officers were offered the training in 305 clusters over 49 months so that although all eventually received the training, there were many time periods in which only some of the officers had been trained – for the study, 22 clusters remained without training throughout the study period. The center panel of Figure 42 shows this design, with blue x's marking months in which a given cluster had already received training and the boundary representing the point where each cluster of officers received training.

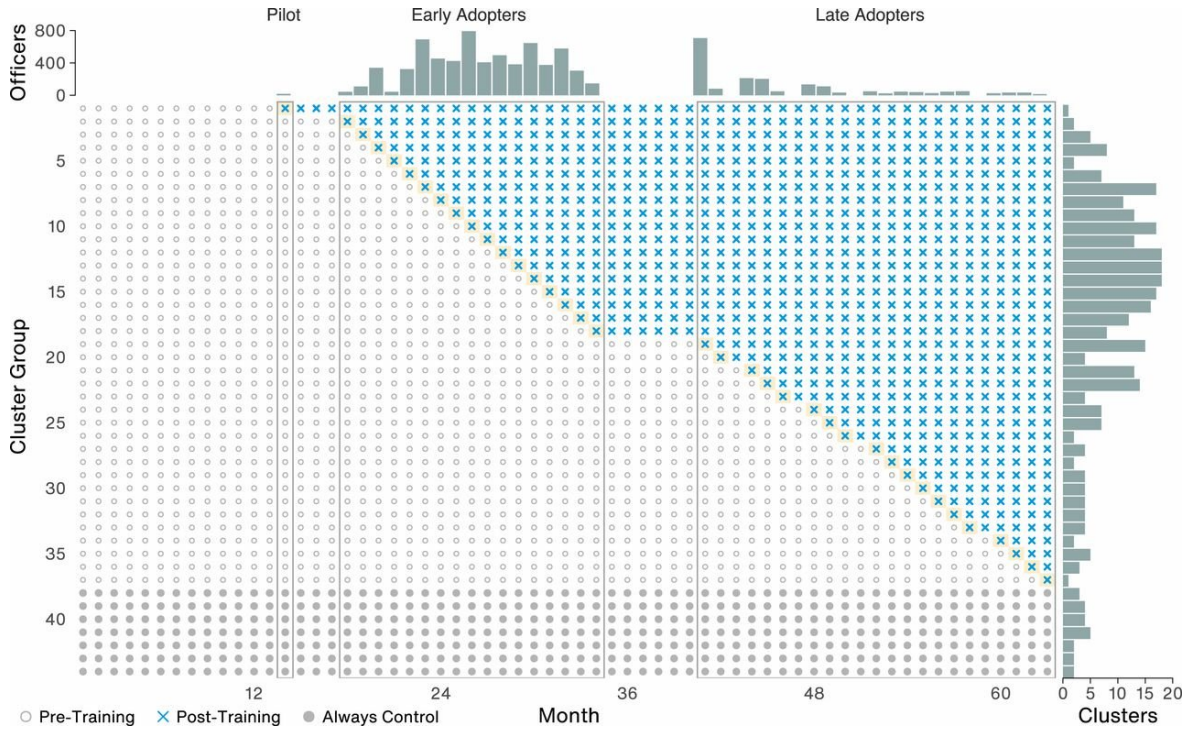


Figure 42: Figure copied from Wood et al. (2020).

To formalize this set-up, we add the following notation:

- Let Z_{it} be the treatment status of unit i at time t
- Let T_i be the time that unit i was first treated with $T_i = \infty$ for any units that are never treated.
- Let $\mathcal{M}_i = \{i' : T_{i'} > T_i\}$ be the set of units treated strictly later than unit i. If unit i is treated, these units serve as controls.
- Let the units be indexed so that $T_1 \leq T_2 \leq ... \leq T_N$. That is, for any $i < j$, unit i was treated earlier or at the same time as unit j and any never-treated units are indexed last.

Alternatively, if we have *groups* of units that are all treated at the same time, we might define

- Treatment groups $g = 1, ..., K$ and a never-treated group 0
- Let G_i be the group ID of unit i
- Let T_g be the time of first treatment for units with $G_i = g$.
- Let the groups be indexed so that $T_1 \leq T_2 \leq ... \leq T_K$. That is, for any $1 \leq g < g' \leq K$, group g was treated earlier than group g' and the group of never-treated units are last with $T_0 = \infty$.

We can approach staggered adoption designs via difference in difference or synthetic control type approaches. A key insight for this kind of set-up is that at each time t, we can use not just the never-treated group but any not-yet-treated units as our control group.

10.5.1 Staggered Adoption and Difference in Differences

For staggered adoption data, we can essentially do a separate DiD estimation for each treatment group g and possibly average them. Let $A = \{i : G_i > g \text{ or } G_i = 0\}$, the set of units treated later than group g or never treated and let $B = \{i : G_i = g\}$, the units in group g. The key DiD assumption is again a **parallel trends** assumption relative to some pre-treatment time $t < T_g$ – usually $T_g - 1$:

$$\mathbb{E}(Y_{iT_g}(0) - Y_{it}(0)|i \in A) = \mathbb{E}(Y_{iT_g}(0) - Y_{it}(0)|i \in B)$$

This says that the on-average trend in the treated group g if it had not been treated is the same as the trend for its control group. Suppose we use $t = T_g - 1$. Given the parallel trends assumption, we can form the usual DiD estimator of the average effect on treated group g as:

$$\widehat{ATT}(g) = \frac{1}{n_A}\sum_{i\in A}(Y_{iT_g} - Y_{i,T_g-1}) - \frac{1}{n_B}\sum_{i\in B}(Y_{iT_g} - Y_{i,T_g-1})$$

where n_A, n_B are the sizes of the treat and control groups. As before, we could also do all of this assuming a parallel trend conditional on some covariates and use associated matching and weighting approaches. Once we pick a pre-treatment and post-treatment time and define groups A, B, we are back in the single DiD world. Given a bunch of $\widehat{ATT}(g)$, we might average them to get an overall average treatment effect on the treated averaged over different treated groups treated at different times. Theoretically, this ATT estimand is

$$ATT = \sum_{g=1}^{K}\Pr(G_i = g)ATT(k) = \sum_{g=1}^{K}\Pr(G_i = g)\mathbb{E}(Y_{iT_g}(1) - Y_{iT_g}(0)|G_i = g)$$

Let n_g be the number of units in group g and n_1 be the total number of treated units. Then the estimator of the above is

$$\widehat{ATT} = \sum_{g=1}^{K}\frac{n_g}{n_1}\widehat{ATT(g)}$$

In principle, we could also identify a average treatment effect at some further out post-treatment time $T_g + c$ and perhaps even average the effect over different time periods. In this case, we would need to assume parallel trend holds also for that time period. However, this becomes less plausible the further the separation between the pre- and post- treatment time period.

10.5.2 Staggered Adoption and Synthetic Control

Ben-Michael et al. (2019) generalize the synthetic control method to the staggered adoption scenario. In this case, we can again, for each unit i (or each collection of units treated at the same time in group g) create a synthetic control using the units that have been untreated up to time T_g. We could again consider averaging over the ATT estimated for each unit and each treated group. But is averaging the best thing we could do? Ben-Michael et al. (2019) distinguish three strategies:

1. The **Unit-by-unit/Separate SCM**[67] is simply the result of doing synthetic control separately for each treated unit and then averaging the results. That is, for each unit i, we calculate weights $w_{ii'}$ for $i' \in \mathcal{M}_i$ based on the observations for the $t = 1, ..., T_i - 1$ time periods before unit i was treated.

$$\arg\min_{w_i}\sum_{t=1}^{T_i-1}\left(Y_{i,T_i-t} - \sum_{i'\in\mathcal{M}_i} w_{ii'}Y_{i',T_j-t}\right)^2 \quad \text{for each} \quad i \tag{31}$$

2. **Pooled approach:** in this approach, we minimize the following pooled measure of imbalance:

$$\arg\min_{w}\sum_{t=1}^{T}\left(\sum_{i:T_i>t}\left(Y_{it} - \sum_{i'\in\mathcal{M}} w_{ii'}Y_{i't})\right)\right)^2 \tag{32}$$

   This is similar to the previous term except notice that the squaring is now on the outside of the sum over $i : T_i > t$. This means that instead of choosing weights to balance to an individual treated unit, we are only balancing an average. We can consider the difference between focal unit i at untreated time t and its synthetic control (weighted averages of unis they are matched to) as a kind of residua. At each time t, we want the residuals for not yet treated units to be small on average, which does not preclude some from being individually large. That is, the squaring being outside the inner sum allows scenarios where some of the individual deviations are large in magnitude but cancel. We do still have a weight $w_{i,i'}$ for each pair of unit i and later-treated unit i'.

3. The main proposal in Eli Ben-Michael and Rothstein (2021) is a **partially pooled** approach which combines the previous two.

Given these weights, we can estimate average treatment effects on the treated for $k = 1, 2, 3...$ periods after treatment onset, with upper limit of k depending on how much data we observe. See Eli Ben-Michael and Rothstein (2021) for details.

[67] SCM stands for Synthetic Control Method

10.6 PanelMatch

The staggered adoption pattern described in the previous section assumes that there is a single time when units are first treated and after that, units are always treated. Of course, in general, units may exhibit more complicated patterns, receiving treatment intermittently. In this case, the basic staggered adoption idea that we might use units that have not yet been treated as controls can be generalized to the idea of looking at units with the same **treatment history** up until a time of interest, thereby controlling for the role of previous treatments. Figure 43 gives a toy example of what this set-up could look like. At times 1 and 2, we could do a simple difference in difference type approach to estimate the average effect on the treated at time 2. However, what if we want to know the average effect on units 4 and 5 at time 3 or on unit 1 at time 4? In this case, we can pattern match. For example, unit 6 has the same treatment history as units 4 and 5 before time 3 and then at time 3, it is control while the others are treated. Similarly, units 2 and 3 have the same treatment history as unit 1 at times $1-3$ and only at time 4 do they diverge. Essentially, we encounter a **matching problem**. As in Module 8, we'd ideally like to exact match on pre-treatment covariates: here, the lagged treatments, though one could also include other covariates. As in Module 8, exact match may not always be possible. Figure 43 gives a real-data example of what this data structure can look like.

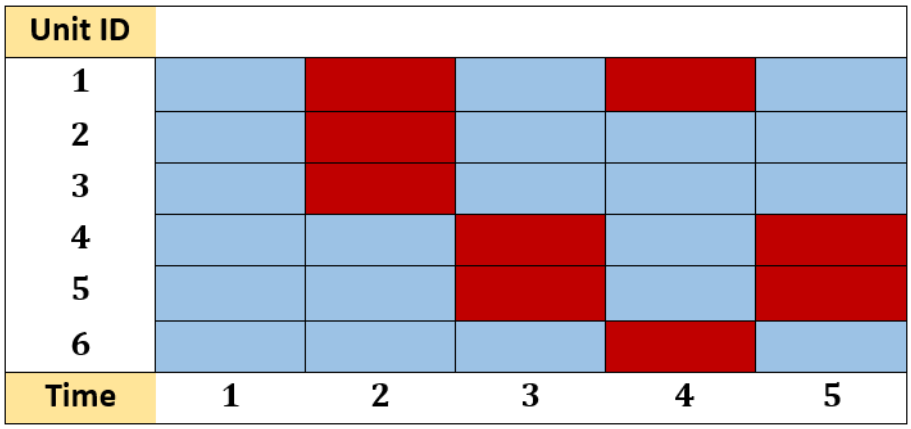


Figure 43: Toy example of more complicated treatment pattern. Darker red blocks represent times when a unit was treated.

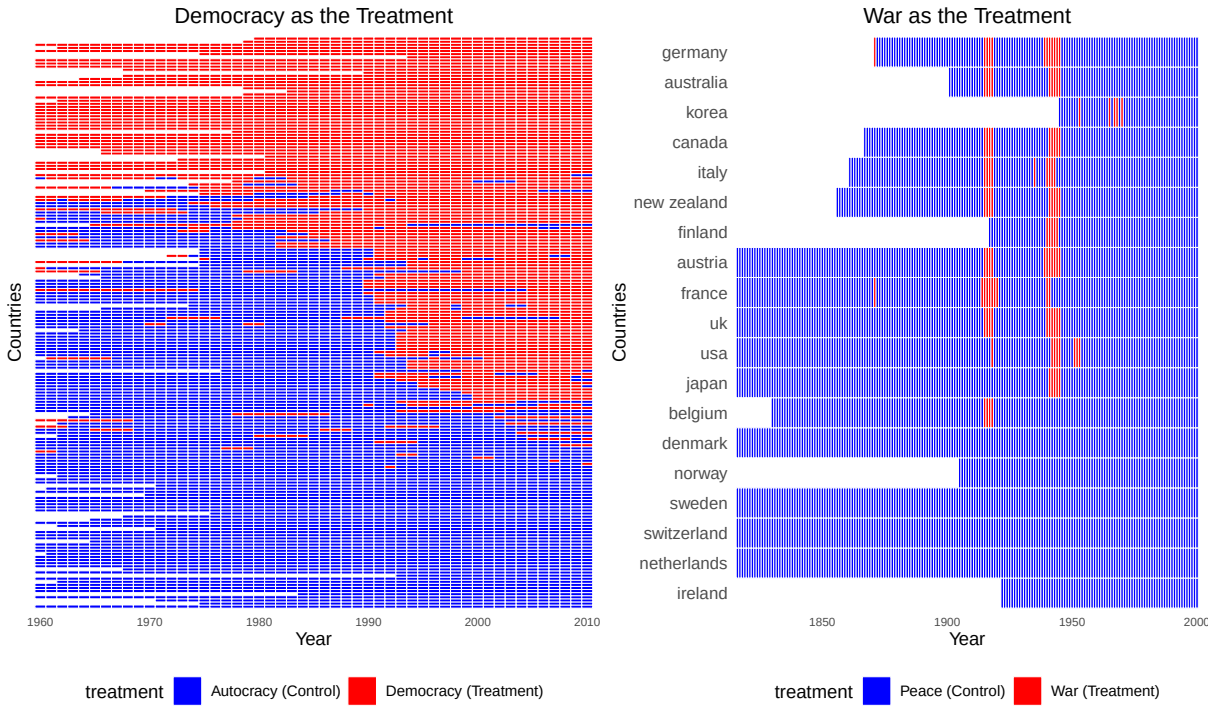


Figure 44: Real examples of more complicated treatment patterns from Imai et al. (2023)

As in the case of staggered adoption, one way we might approach a data structure like this is to do lots of mini Difference-in-Difference type comparisons, isolating groups of units which were all untreated at time t, with some treated at later time t'. Alternatively, we might, as already discussed in Section 10.3, fit some regression model, such as the two-way fixed effects regression:

$$Y_{it} = \alpha_i + \gamma_t + \beta X_{it} + \epsilon_{it}$$

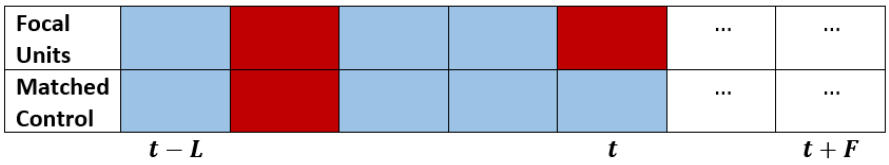
where α_i is a unit-specific effect, γ_t a time-specific effect, and β a treatment effect. However, while in Section 10.1.2, we showed that this regression model is actually equivalent to DiD in the 2×2 DiD case, this equivalence is not true in general.

10.6.1 Set-up and Estimand

Returning to the 'lots of mini-DiD's' idea, Imai et al. (2023) introduce a method and `R` package they call PanelMatch which involves matching on treatment histories as described for the toy example above. Formally, we define:

- L - the number of treatment history of 'lag' periods to consider when identifying matches
- F the number of future time periods after treatment ('lead' periods) to jump ahead to when estimating the treatment effect
- T the total number of time periods observed
- Y_{it} the observed outcome at time t
- X_{it} the treatment status at time t
- Potential outcome at time $t+F$, $Y_{it+F}(x_{i,1:t}) = Y_{i,t+F}(x_{i1}, ..., x_{it})$, which is a function of the entire treatment history up to time t (here we are essentially ignoring treatments at times $t+1, ..., t+F-1$, though these could implicitly be part of the pathway that determines what this potential outcome is).

The **estimand** is the average treatment effect on the treated at time $t+F$ for those treated in time period t and untreated in time period $t-1$. We get this by matching to units who had the same treatment pattern for times $t-L, ..., t-1$ but were untreated at time t.



Formally, the estimand is

$$ATT = \mathbb{E}[Y_{i,t+F}(X_{it}=1, X_{i,t-1}=0, \{X_{i,t-l}\}_{l=2}^{L}) \; - \; Y_{i,t+F}(X_{it}=0, X_{i,t-1}=0, \{X_{i,t-l}\}_{l=2}^{L}) \quad | \quad X_{it}=1, X_{i,t-1}=0] \quad (33)$$

This is called the **Average Treatment Effect of Policy Change for the Treated**. There is a bit of notational ambiguity here where it looks as if this might be the ATT at a particular time t but it is meant to be the ATT over all 'policy change' instances occurring at any time. That is, if we take all the instances over all times and units where a unit goes from being untreated at time $t-1$ to being treated at time t, we are estimating the average effect of this.[68] One nuance here is that if we require a lag period of size L, then we are only averaging over policy changes at times $t \geq L+1$ and if we set a lead of F, the potential outcomes above are only defined up to $t = T - F$. This is reflected in our estimator. For the estimand in Equation 33, the left side is something we can immediately identify from our observations. The right side is where we need matching on treatment histories.

[68] One could imagine other estimands as well, such as the effect of *first* treatment – that would be more the staggered adoption idea.

10.6.2 Estimation Procedure

The Estimation Procedure for the average effect above is as follows:

1. Find the policy change instances: treated units i at time t with $X_{i,t-1} = 0, X_{it} = 1$
2. For each such instances, form a matched set M_{it} of control observations with identical treatment histories in the lag period $[t - L, t - 1]$. [69]
3. Optionally, refine the match groups via matching and weighting techniques. This could involve throwing out units from M_{it} if their covariates are too different or weighting the units within the match group to achieve covariate balance relative to the treated units with the same history. If we do not do this step, the default weights are $w_i^{i'} = \frac{1}{|M_{it}|}$ for units $i' \in M_{it}$, which gives an average in the estimator below.
4. **Average over difference in differences estimators:**

$$\widehat{ATT} = \frac{1}{\sum_{i=1}^{N}\sum_{t=L+1}^{T-F} D_{it}} \sum_{i=1}^{N}\sum_{t=L+1}^{T-F} D_{it} \left((Y_{i,t+F} - Y_{i,t-1}) - \sum_{i' \in M_{it}} w_i^{i'} (Y_{i',t+F} - Y_{i',t-1}) \right)$$

where $D_{it} = I(X_{it} = 1, X_{i,t-1} = 0)$, the indicator for whether a unit went from being untreated to treated at time t and $w_i^{i'}$. Notice this is exactly a difference in differences estimator calculated using the matched group of units that was untreated at time $t - 1$ and stayed untreated at time t. The key assumption is **parallel trends** for each of these DiD estimations, conditional on the pre-treatment history:

$$\mathbb{E}(Y_{i,t+F}(0) - Y_{i,t-1}(0) | X_{i,t-1}, ..., X_{i,t-L}, X_{i,t} = 1) = \mathbb{E}(Y_{i,t+F}(0) - Y_{i,t-1}(0) | X_{i,t-1}, ..., X_{i,t-L}, X_{i,t} = 0)$$

As before, this parallel trends assumption is fundamentally unverifiable, but we can evaluate its plausibility by looking at the control and match groups' outcomes over the pre-treatment period and evaluating whether they appear parallel. See Figure 5 in Imai et al. (2023) for an example.

[69] Note that especially if the lag period is too large, there may sometimes not be any exact matches.

11 Conclusion

Causal inference formalizes many intuitive notions about how to tell whether a treatment 'makes a difference.' At the same time, it reveals implicit and sometimes less intuitive assumptions needed for observed statistics to have causal interpretations and gives us a language for what can go wrong. We hope you have by now internalized the distinctions between defining causal estimands, identifying them, estimating them, and quantifying the uncertainty in those estimates. These distinctions are fundamental and too often overlooked. Although the first two are sometimes seen as 'causal inference proper' and the latter two as 'statistics,' the whole sequence is really fundamental to what it means to do causal inference in practice. Even when, at the end of the day, causal inference concludes by running regressions, the justifications and assumptions underlying those regressions are nuanced and critical for science and policy-making. Correlation is not causation...but sometimes, if you're careful, it is. And via potential outcomes, we can now say what we mean by causality in the first place.

In this course, we started with the language of potential outcomes and used it to describe the world of experiments. There, by randomly assigning treatment – or at least an encouragement to be treated – we hope that other factors 'balance' on average. With a few important exceptions, randomization has stayed with us throughout as one of the fundamental principles for identifying causal effects. When we moved to observational studies, conditional randomness or unconfoundedness became the key assumption, the idea being that at least within groups with similar baseline characteristics, who receives treatment might be essentially random. Given unconfoundedness, we explored both outcome modeling and matching/weighting approaches to identify average causal effects as well as doubly robust estimators, which combine them. We briefly looked at causal mediators, which are often of substantive interest but require still more tricky unconfoundedness assumptions.

Yet assuming unconfoundedness is not enough, as it is often not entirely plausible. Here, we briefly introduced you to sensitivity analysis and partial identification, which examine how bad confounding dynamics would have to be to fully account for apparent causal effects or try to bound a causal quantity of interest at the most extreme values it could possibly taken given what we can observe. It is easy to list some assumptions and run some calculations, but in messy reality, confronting issues like selection bias and evaluating sensitivity to assumptions is crucial.

The exceptions to the randomization rule were regression discontinuity designs, difference in difference, and synthetic control methods. These methods instead relied on forms of principled extrapolation: continuity of potential outcome curves on either side of a cut-off or parallel trends over time. These methods are highly intuitive, but again, potential outcomes help us think more explicitly about what we are assuming and what could go wrong.

There are many more areas of causal inference that we did not cover, including (but not limited to):

- Complex experimental designs
- Experiments in the context of networks, accounting for complex spillover effects
- Sequential experiments and reinforcement learning
- Causal machine learning
- Individualized treatment rules and policy learning
- The do-calculus and further DAG-based approaches
- Causal discovery
- Proximal causal inference
- Causal inference with unstructured data (text, video, sound)
- Causal definitions of fairness / connections to the literature on algorithmic fairness

Many of these are active areas of current research. This course was just the gateway!

A A few key results used in this course

1. **Law of Total Probability**: given event A and events $B_1, ..., B_k$ where B_i's are mutually exclusive (no two can happen at same time) and exhaustive (account for all possibilities), we have

$$P(A) = \sum_{i=1}^{k} P(A, B_i) = \sum_{i=1}^{k} P(A|B_i)P(B_i)$$

Sometimes A is a big, complicated event that it is hard to calculate the probability for but once we condition on B_i, it simplifies things and makes calculations much easier!

2. **The Law of Total Expectation**[70] says that for random variables X and Y,

$$\mathbb{E}(\mathbb{E}(Y|X)) = \mathbb{E}(Y)$$

This also holds conditionally. For example, below we condition on Z throughout

$$\mathbb{E}(\mathbb{E}(Y|X, Z)|Z) = \mathbb{E}(Y|Z)$$

This result is EXTREMELY useful for dealing with situations where we have multiple sources of randomness. Figuring out $\mathbb{E}(XY)$ might be hard. Figuring out $\mathbb{E}(\mathbb{E}(XY|X)) = \mathbb{E}(X\mathbb{E}(Y|X))$ might be easier because we can deal first with randomness in Y given fixed X and then with randomness in X!

If X is a categorical variable (e.g., suppose it is binary), it can also sometimes be very useful to write out

$$\mathbb{E}(Y) = \mathbb{E}(\mathbb{E}(Y|X)) = \mathbb{E}(Y|X = 1)P(X = 1) + \mathbb{E}(Y|X = 0)P(X = 0)$$

and deal with each of these four parts separately!

3. **The Law of Total Variance** says that

$$\mathbb{V}(Y) = \mathbb{V}(\mathbb{E}(Y|X)) + \mathbb{E}(\mathbb{V}(Y|X))$$

For intuition, imagine X is binary. $\mathbb{V}(\mathbb{E}(Y|X))$ characterizes how much the expected value of Y varies between $X = 0$ and $X = 1$ (across-group variation) while $\mathbb{E}(\mathbb{V}(Y|X))$ characterizes the average variance within the $X = 0$ and $X = 1$ groups (within-group variation)

[70] Also known as the Law of Iterated Expectation or, if you've taken Stat 210, Adam's Law.

B Some things to know about linear regression for Stat286/Gov2003

For this course, you do not need to know all the in and outs of regression, but you should be comfortable with the basic set-up for regressing a vector of observations Y on some set of covariates X. The following is mostly meant as a recap, but some details may be new to you.

B.1 The Model

The multivariate linear model is:

$$Y_i = \beta_0 + X_i^T \beta + \epsilon_i$$

where $X_i \in \mathbb{R}^p$ is a vector of observed covariates, $y_i \in \mathbb{R}$ is an observed response, $\beta_0 \in \mathbb{R}$ is an unknown intercept term, $\beta \in \mathbb{R}^p$ is the unknown parameter characterizing how Y_i changes with x_i and ϵ_i is a random error term. Note that X_i and β are length p column vectors here so $X_i^T \beta$ is the dot product:

$$X_i^T \beta = X_1 \beta_1 + ... + X_p \beta_p$$

In full matrix-vector notation, the regression for all $i = 1, ..., n$ observations is often written

$$Y = X\beta + \epsilon$$

where here $X \in \mathbb{R}^{n \times p+1}$ is a $n \times p+1$ matrix such that the first column is a column of 1's for the intercept columns and the remaining columsn contain the data for each covariate. Each row contains the observations for a single unit across all covariates. In this notation, the intercept is treated as part of the $\beta \in \mathbb{R}^{p+1}$. $Y \in \mathbb{R}^n$ is a $n \times 1$ vector of outcome variables for each observation, and ϵ is a $n \times 1$ vector of error terms for each observation. Note that some textbooks may use $p' = p + 1$, treating the intercept term as part of the count of covariates.

The **standard assumptions** made about this model are as follows, though note that not all of them are necessary for all results below.

1. **Linearity**: model is correctly specified so that the conditional mean is actually of the form $X\beta$ (linear in parameter β)
2. **Exogeneity**: $E(\epsilon_i|X) = 0$, which implies
   (a) $E(\epsilon_i) = E(E(\epsilon_i|X)) = 0$
   (b) $E(\epsilon_i x_{jk}) = E(x_{jk} E(\epsilon_i|X)) = 0$ for all $j = 1, ..., n, k = 1, ..., p$, $i = 1, ...n$
   (c) $Cov(\epsilon_i, x_{jk}) = E(\epsilon_i x_{jk}) - E(\epsilon_i)E(x_{jk}) = 0$. For $i \neq j$, independent observations is enough to imply this. For $i = j$, $Cov(\epsilon_i, x_{ik}) = 0$ says that an observation's error is not correlated with its value of covariate k (violations may be diagnosed by a pattern in the residuals).

   Intuition: says x not predictive of the error term in a first order (mean) way. Full $X \perp \epsilon$ is sufficient but not necessary here as we can have exogeneity but not homoskedasticity.
3. **Homoskedasticity**: $Var(\epsilon_i|X) = \sigma^2$ for all i.
   Intuition: x is not predictive of error term in second order (variance) way (may be diagnosed by residuals being systematically greater or smaller as x varies).
4. **Uncorrelated errors** $Cov(\epsilon_i, \epsilon_j) = 0$. This would be true, for example, if we have an i.i.d. sample. Together with homoskedasticity, this implies the variance-covariance matrix of the vector ϵ is $Var(\epsilon) = \sigma^2 I_n$
5. **Rank condition:** no perfect linear dependence in columns of X so rank$(X) = p$
6. **Normality** $\epsilon_i \sim N(0, \sigma^2)$

B.2 Estimation

The least squares estimator of β minimizes the sum of squared deviations from $X_i^T\beta$ and has the following form:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - X_i^T\beta)^2 = (X^TX)^{-1}X^TY$$

This is well-defined given Assumption 5, which makes (X^TX) invertible. If well-defined, only Assumptions 1 and 2 are necessary for $\hat{\beta}$ to be **unbiased**. The proof of unbiasedness is:

$$\mathbb{E}(\hat{\beta}|X) = (X^TX)^{-1}X^TE(Y|X) = (X^TX)^{-1}X^TE(X^T\beta + \epsilon|X) = \beta$$

(note: for finite-sample results, we generally treat X as fixed throughout and all results are conditional on it)

Given the estimator $\hat{\beta}$, we can calculate **fitted values** (estimates of $E(Y|X = x_i)$) as $\hat{y}_i = x_i^T\beta$. The **residual** or leftover variation in y_i that isn't accounted for by the fitted value, is defined as $r_i = y_i - \hat{y}_i$. It is often convenient to put these into vectors:

$$\hat{Y} = X\hat{\beta} \qquad R = Y - \hat{Y}$$

In statistics, we are usually interested in some kind of uncertainty quantification. For this, we need to be able to estimate the variance in Y_i's. Under Assumptions 1-6, this variance is homoskedastic and we have the following unbiased estimator for the variance σ^2 of Y

$$\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n} r_i^2 = \frac{R^TR}{n-p} = \frac{Y^T(I-H)Y}{n-p}$$

Under the same assumptions, we can also consider the variance of our estimator $\hat{\beta}$ as well as of the fitted values $\hat{Y}$ and residuals R. The following results can be derived from the form of $\hat{\beta}$ and the rules for taking expectations and variances of linear functions of random vectors.[71]

$$Var(\hat{\beta}|X) = \sigma^2(X^TX)^{-1} \qquad Var(\hat{Y}|X) = \sigma^2H \qquad Var(R|X) = \sigma^2(I-H)$$

Where $H = X(X^TX)^{-1}X^T$ is the projection matrix (See below). We can estimate these by plugging in $\hat{\sigma}^2$

B.2.1 Special case: Categorical Covariates

In general, it may be very messy to write down the value of a particular $\hat{\beta}_j$ in closed form because of the inversion of $(X^TX)^{-1}$. However, in the case of a model with just categorical covariates, we can write these down in closed form because each is just some combination of averages. In regression, categorical covariates are dummy-encoded into columns of 0's and 1's, each encoding a level of the variable. Usually, this is parameterized so that one variable is chosen as a reference level and encoded via an intercept column of 1's. This is also the default behavior of functions such as `lm()` in `R`.

For example, suppose I have 6 observations of weight in kilograms Y for some animals. I have a variable X that takes the 3 levels, "cat", "dog", "iguana," and 2 observations from each of these levels. Suppose I use cat as the reference level. Then the regression model is:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_2 \\ Y_4 \\ Y_5 \\ Y_6 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_2 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \end{bmatrix}$$

For reference level observations (cat), we have $Y_i = \beta_0 + \epsilon_i$. For non-reference levels, we have for dog: $Y_i = \beta_0 + \beta_1 + \epsilon_i$ and for iguana, $Y_i = \beta_0 + \beta_2 + \epsilon_i$. When we regress Y on X, the intercept estimate $\hat{\alpha}$ will be the sample mean of the reference level (mean weight of cats $\bar{Y}_1$) while the $\hat{\beta}_j$ for each of the other levels j will be the deviations $\bar{Y}_j - \bar{Y}_1$ from the reference level (since then the fitted value is $\bar{Y}_1 + \bar{Y}_j - \bar{Y}_1 = \bar{Y}_j$). Overall, this regression gives the sample mean for each level. Notice that this does not make any linearity assumptions – we are not assuming any particular relationship between the weights of cats, dogs, and iguanas.

Proof: There are multiple ways to prove that the regression coefficient with one categorical predictor returns just the sample means for the predicted values. One, which I show below is to simply take the derivative of the least squares condition we want to minimize to get the following condition

[71] The key rules: for constant vector b, random vector W, and matrix A, $E(AW + b) = AE(W) + b$ and $Cov(AW + b) = ACov(W)A^T$ where Cov denotes the variance-covariance matrix.

$$\frac{\partial}{\partial \beta_j} \sum_{i=1}^{n} (Y_i - X_i^T \hat{\beta})^2 = 2 \sum_{i=1}^{n} X_{ij}(Y_i - X_i^T \hat{\beta}) = 0$$

where in this case, I let $j = 1, ..., p$ with $j = 1$ marking the intercept column. That is, we have a categorical predictor with p levels and the first is the reference level. Recall that because of the dummy encoding, $X_{ij} \in \{0, 1\}$ and for a unit such that $X_{ij} = 1$, we have $X_i = (1, 0, ..., 0, 1, 0, ..., 0)$ – that is, with a 1 only appearing in the intercept position and the j^{th} position. Hence, for $j > 1$, the condition above is equivalently

$$\sum_{i:X_{ij}=1} (Y_i - \hat{\beta}_1 - \hat{\beta}_j) = 0 \quad \leftrightarrow \quad \hat{\beta}_1 + \hat{\beta}_j = \frac{1}{n_j} \sum_{i:X_{ij}=1} Y_i = \bar{Y}_j$$

where n_j is the number of units coming from the level that β_j corresponds to and $\bar{Y}_j$ is the average outcome for that group. Finally, consider the case of $j = 1$. We then have

$$\begin{aligned}
\sum_{i=1}^{n} Y_i &= (\sum_{i=1}^{n} X_i)^T \hat{\beta} \\
\sum_{i=1}^{n} Y_i &= \begin{bmatrix} n & n_2 & ... & n_p \end{bmatrix} \hat{\beta} \\
\sum_{i=1}^{n} Y_i &= n\hat{\beta}_1 + n_2(\bar{Y}_2 - \hat{\beta}_2) + ... + n_p(\bar{Y}_p - \hat{\beta}_p) \\
\sum_{i=1}^{n} Y_i &= n\hat{\beta}_1 - (n - n_1)\hat{\beta}_1 + \sum_{j=2}^{p} \sum_{i:X_{ij}=1} Y_i \\
\sum_{i=1}^{n} Y_i - \sum_{j=2}^{p} \sum_{i:X_{ij}=1} Y_i &= n_1 \hat{\beta}_1 \\
\bar{Y}_1 &= \hat{\beta}_1
\end{aligned}$$

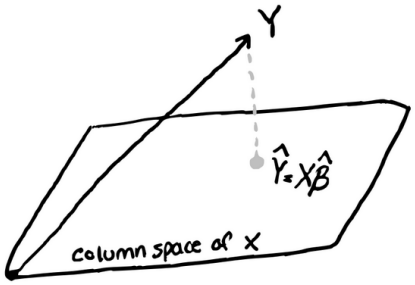
That is, the intercept is the mean for the reference group and the other $\hat{\beta}_j$'s are $\hat{Y}_j - \hat{Y}_1$.

B.2.2 Hypothesis Testing and Confidence Intervals

Normality (or approximate normality) is necessary for standard hypothesis tests and confidence intervals (multiplying standard error by 2) to be valid. For example, given the normality assumption that $Y \sim N(0, \sigma^2 I_n)$, we can argue that $\hat{\beta} \sim N(\beta, \sigma^2 (X^T X)^{-1})$ and use this to justify hypothesis tests that compare to the normal (or, given we need to estimate σ^2, the t-distribution). The exact details for hypothesis testing for linear models are not given here.

B.3 Further Useful Results and Perspectives

B.3.1 Regression as Projection



OLS regression can be viewed as the projection of vector Y onto the space spanned by the columns of matrix X using projection matrix $H = X(X^TX)^{-1}X^T$.[72] We then have that vectors of fitted values $\hat{Y}_i$ and residuals $r_i = Y_i - \hat{Y}_i$ can be written:

$$\hat{Y} = X\hat{\beta} = HY \qquad\qquad R = Y - \hat{Y} = (I - H)Y$$

Note: the projection matrix H is idempotent $H^2 = H$ and symmetric $H = H^T$. This matrix is used in Module 4.

B.3.2 Connection between β and Covariances (univariate case)

Let X and Y be univariate random variables with $Y = X\beta + \epsilon$ and assume $Cov(X, \epsilon) = 0$ (exogeneity). Then

$$\text{Cov}(X, Y) = \text{Cov}(X, X\beta + \epsilon) = \text{Cov}(X, X)\beta + \text{Cov}(X, \epsilon) = \text{Cov}(X, X)\beta$$

solving for β then gives, assuming the covarinace matrix of X is invertible,

$$\beta = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}$$

Note that we also have $\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$, so the estimator is a ratio of empirical covariances (the $\frac{1}{n}$ cancel). Intuitively, this tells us $\hat{\beta}$ is measuring how Y changes with X, relative to the spread of X. There are generalizations of this beyond the univariate case.

B.3.3 R^2

One way of measuring the quality of a regression is R^2, which captures the fraction of variation in Y that is accounted for (in a correlation sense) by variation in X. $TSS = \sum_{i=1}^n (Y_i - \bar{Y}_i)^2$ be the total sum of squares, $RSS = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$ be the residual sum of squares and $ESS = \sum_{i=1}^n (\hat{Y}_i - \bar{Y}_i)^2$ be the explained sum of squares. Then we have the decomposition

$$TSS = ESS + RSS$$

and R^2 is defined as

$$R^2 = 1 - \frac{RSS}{TSS} = \frac{ESS}{TSS}$$

In the simple linear regression case (again analogues exist for multivariate regression), the observational R^2 above is really an

[72] Note: I use H instead of the P used in lecture slides since P is so prevalent in the notation already.

estimator of the following theoretical quantity.

$$\begin{aligned}
R^2 = \frac{V(\hat{Y})}{V(Y)} &= \frac{V(\hat{Y})V(\hat{Y})}{V(Y)Var(\hat{Y})} \\
&= \frac{(Cov(\hat{Y},\hat{Y}) + Cov(\hat{Y},\epsilon))^2}{V(Y)Var(\hat{Y})} \\
&= \frac{(Cov(\hat{Y}+\epsilon,\hat{Y})^2}{V(Y)Var(\hat{Y})} \\
&= \frac{(Cov(Y,\hat{Y})^2}{V(Y)Var(\hat{Y})} \\
&= \frac{(Cov(Y,X)^2\beta^2}{V(Y)Var(X_i)\beta^2} \\
&= \frac{(Cov(Y,X)^2}{V(Y)Var(X)} \\
&= Corr^2(Y,X)
\end{aligned}$$

Here we are using the fact that $Cov(\hat{Y},\epsilon) = \beta^T Cov(X,\epsilon) = 0$ by the exogeneity assumption. Overall:

$$R^2 = corr(Y,\hat{Y})^2 = corr(Y,X)^2$$

This will also hold if we plug in the *sample* covariances between these quantities.

Notice also: this implies that in the univariate case, R^2 it the same for Y regressed on X and X regressed on Y (even though other regression quantities are not per se the same!)

B.4 Frisch-Waugh-Lovell Theorem

The FWL provides some insight into what a multivariate regression is actually doing and can be a useful decomposition for proving things. At a high level, it says that given regression $Y = X\beta + Z\gamma + \epsilon$, I can break this into two steps: (1) regressing Y and X on Z and (2) regressing variation leftover in Y on variation leftover in X.

Theorem: Let X and Z be $n \times k$ and $n \times p$ matrices of covariates respectively. Consider the linear regression model $Y = X\beta + Z\gamma + \epsilon$, for which the standard OLS estimator for regressing Y on $W = [X, Z]$ is $(\hat{\beta},\hat{\gamma})^T = (W^TW)^{-1}W^TY$. Consider taking the following three steps:

1. Regress Y on Z to obtain $\hat{\delta}_1 = (Z^TZ)^{-1}ZY$ and calculate $n \times 1$ residual vector $r_y = Y - Z\hat{\delta}$
2. For each column j of X, regress X_j on Z to obtain $\hat{\delta}_{2j} = (Z^TZ)^{-1}Z^TX_j$ and calculate residual vector $r_{x_j} = X_j - Z\hat{\delta}_{2j}$. Form matrix R_X as $R_X = [r_{x_1}, ..., r_{x_k}]$
3. Regress residuals from (1) on residuals from (2) to form $\hat{\delta}_3 = (R_X^TR_X)^{-1}R_X^Tr_y$

Then $\hat{\delta}_3 = \hat{\beta}$. That is, the regression result from the steps above is algebraically equivalent to what we would get for β from regressing Y on $W = [X, Z]$ immediately. Moreover, $\hat{\delta}_1 - \hat{\delta}_2\hat{\delta}_3 = \hat{\gamma}$

A succinct way of stating this result in terms of projection matrices is as follows: let $M^\perp = I - Z(Z^TZ)^{-1}Z^T$ be the projection matrix for projecting on the space orthogonal to col(Z). Then the OLS estimate of β from regressing Y on $[X, Z]$ is equivalent to the OLS estimate for the following regression:

$$M^\perp Y = M^\perp X\delta_3 + M^\perp\epsilon$$

That is, $\hat{\beta} = \hat{\delta}_3 = (X^TM^\perp X)^{-1}X^TM^\perp Y$

Intuition: It is easier to gain some intuition for categorical Z. Imagine that Z corresponds to treatment (0 or 1), Y is blood pressure, and X is age. Suppose that the treatment lowers blood pressure so that within the $Z = 1$ group, Y is lower on average. Suppose also that treatment is correlated with age with older patients more likely to receive treatment. The FWL Theorem says that regressing Y on age and treatment is equivalent to first accounting for treatment by (1) regressing blood pressure on treatment and (2) regressing age on treatment and then (3) looking at how the remaining variation in age around the mean age for each treatment group correlates with remaining variation in blood pressure. If, for example, units with higher-than-average-within-treatment-group-age have higher-than-average-within-treatment-group blood pressure, then we get a larger effect of age, holding treatment Z constant (which is the usual way of interpreting multivariate regression coefficients!). Figure 45 illustrates the idea. Note however that despite talk of treatments in this example, this is a purely algebreic, correlation-based result. We could equivalently first regress Y on X and and Z on X.

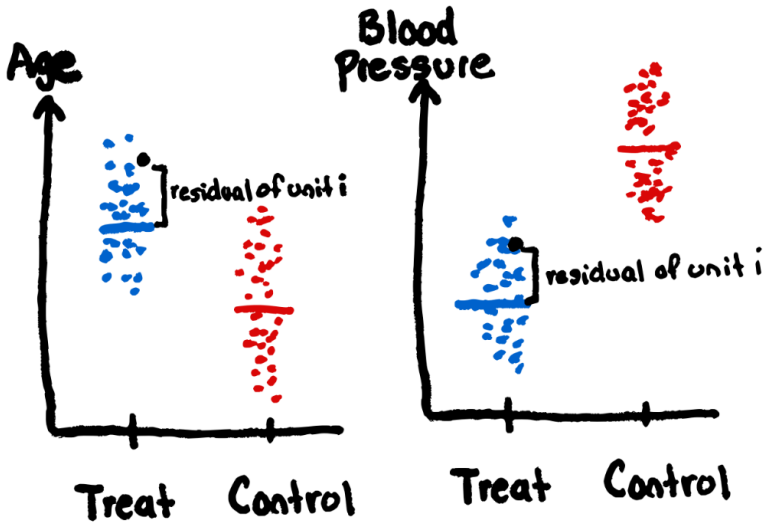


Figure 45: Illustrative Example of FWL idea

Further Resources

Textbooks on linear regression include: Agresti (2015) (Chapter 2 mainly relevant for this course, very statistical, graduate level) and Rawlings (1998) (Chapters 1-4 and 9 most relevant, especially 3, this book much more accessible). A good online resource is https://mattblackwell.github.io/gov2002-book/06_linear_model.html (Chapter 5-7, with set-up and notation that nicely connects to causal inference). There are multitudes of other textbooks as well as videos on YouTube if you need a refresher (e.g., this one gives a good overview of standard OLS assumptions https://www.youtube.com/watch?v=a1ntCyeoJ0k).

C Things to know about regression - Part II: Violations of standard assumptions

This section is less essential knowledge for stat 286 but heteroskedasticity in particular does come up and the other sections give some perspective on key assumptions.

C.1 Linearity and Misspecification (A1)

What does it mean for linearity to be violated so that the linear model is mis-specified? If it is, what does the 'true value of β' even mean?

First, an example of the model not being linear would be if the true model is $Y_i + X_i^2\gamma + \eta_i$ but we fit the model $Y_i = X_i\beta + \epsilon_i$. Our model is then said to be **mis-specified**. In this case, even if all other assumptions hold, the OLS estimator is biased since, letting $X = (X_1, ..., X_n)^T$ and $Z = (X_1^2, ..., X_n^2)$:

$$E(\hat{\beta}|X) = (X^TX)^{-1}X^TE(Y|X) = (X^TX)^{-1}X^TZ\gamma \neq \beta$$

However, this issue is relatively easily fixable. If we regress Y_i on $Z_i = X_i^2$, we are back to linearity. A more difficult case would be if the true model is $Y_i = e^{\beta X} + \eta_i$ and we fit $Y_i = X_i\beta + \epsilon_i$. In this case, there is no function f such that $Y_i = f(X_i)\beta + \epsilon_i$ gets us back to the true model (we would need GLMs).

Second,[73] even if the model is mis-specified, there still exists a **best linear predictor** of the random variable Y given the random variables $X = (X_1, ..., X_p)^T$. In particular, the best linear predictor is defined as vector β that yields the linear combination $X_1\beta_1 + ... + X_p\beta_p$ most highly correlated with Y:

$$\beta = \arg\max_{c \in \mathbb{R}^p} \text{Corr}(Y, X^Tc)$$

This β can be derived in closed form in terms of the covariances between Y and X to be $\text{Cov}(X,X)^{-1}\text{Cov}(X,Y)$. The key point here is that it exists **regardless** of whether the model is truly linear! It may be a very **bad** predictor, but it is still the best *linear* predictor and that is what linear regression using n observations of $(Y, X_1, ..., X_p)$ approximates.

C.2 Exogeneity and Omitted Variable Bias (A2)

Suppose the true model is $Y_i = \xi + X_i\delta + U_i\gamma + \eta_i$ and suppose it is true that $E(\eta_i|X, U) = 0$ (this also implies $E(\eta_i|X) = 0$ - can you reason why?). Suppose U_i and X_i are **correlated** so that $E(U_i|X_i) \neq E(U_i)$. However, we pose the model $Y_i = \alpha + X_i\beta + \epsilon_i$. Here, it is true that Y_i changes linearly with X_i *but* we do not have exogeneity because our error term ϵ_i is

$$\epsilon_i = Y_i - (\alpha - X_i\beta) = (\xi + X_i\delta + U_i\gamma + \eta_i) - (\alpha - X_i\beta) = (\xi - \alpha) + X_i(\delta - \beta) + U_i\gamma + \eta_i$$

Meaning that:

$$\begin{aligned} E(\epsilon_i|X) &= (\xi - \alpha) + X_i(\delta - \beta) + E(U_i|X)\gamma + E(\eta_i|X) \\ &= (\xi - \alpha) + X_i(\delta - \beta) + E(U_i|X)\gamma \neq 0 \end{aligned}$$

What if $E(U_i|X_i) = E(U_i)$ as would be the case if X_i and U_i are independent? Then it might look like we then still get a non-zero expectation, but in fact, in this case, the U_i can safely be added to the error term. If we rewrite the true model as follows using a $\pm E(U_i)\gamma$ trick:

$$Y_i = (\xi + E(U_i)\gamma) + X_i\delta + (U_i - E(U_i))\gamma + \eta_i$$

then, letting $\alpha = \xi + E(U_i)\gamma$ and $\beta = \delta$ and $\epsilon_i = (U_i - E(U_i))\gamma + \eta_i$, we have

$$Y_i = \alpha + X_i\beta + \epsilon_i$$

with

$$E(\epsilon_i|X) = E(U_i - E(U_i)|X)\gamma + E(\eta_i|X) = (E(U_i) - E(U_i))\gamma + 0 = 0$$

it still might be the case that U_i explains some of the variation in Y_i and that not including the U_i leads to a higher variance σ^2, but we do have exogeneity and hence $\hat{\beta}$ will be unbiased for β.

[73] This part is definitely not content you are responsible for in stat 286 / gov 2003

C.2.1 Bias Formula

Putting the above in matrix notation, suppose the true model is $Y = X\beta + U\delta + \eta$ where $E(\eta|X,U) = 0$ and we fit the model $Y = X\beta + \epsilon$. If we compute our standard $\hat{\beta}$ using regression of Y only on X in the set-up above, then

$$\begin{aligned} E(\hat{\beta}|X,U) &= (X^TX)^{-1}X^TE(Y|X,U) \\ &= (X^TX)^{-1}X^T(X\beta + U\delta) \\ &= \beta + (X^TX)^{-1}X^TU\delta \end{aligned}$$

and hence the omitted variable bias conditional on X and U is

$$E(\hat{\beta}|X,U) - \beta = (X^TX)^{-1}X^TU\delta$$

Notice here that $(X^TX)^{-1}X^TU$ is exactly the coefficient we would get if we regressed U on X. It represents the correlation between X and U. If X and U are orthogonal, the bias is 0. Suppose we now take the expectation over U. We then have an expectation of the unbiased estimator for the regression of U on X. That is if we have $E(U|X) = X\gamma$, we get

$$E(\hat{\beta}|X) - \beta = E((X^TX)^{-1}X^TU|X)\delta = \gamma\delta$$

Formulating this in terms of covariance for univariate U,X: using the result from Section B.3.2 that $\hat{\gamma}$ is the ratio of sample covariances (denote these with a hat) and γ can be formulated as a ratio of theoretical covariances:

$$\begin{aligned} E(\hat{\beta}|X,U) - \beta &= \frac{\widehat{Cov}(X,U)}{\widehat{Var}(X)}\delta \\ E(\hat{\beta}|X) - \beta &= \frac{\text{Cov}(X,U)}{\text{Var}(X)}\delta \end{aligned}$$

Hence if there is no association between Y and U (so $\delta = 0$) or X and U are uncorrelated, then there is no bias.

Extension using FWL Theorem: What if instead, I have true model $Y = X\gamma + T\beta + U\delta + \eta$ but fit only $Y = X\gamma + T\beta + \eta$? That is, I have two observed covariates. Perhaps the γ are nuisance parameters that I am not really interested in and what I really want to know is β and in particular, the omitted variable bias for β. Here, we can make use of the **Frisch-Waugh-Lovell Theorem** (Section B.4) to first regress each of Y, T, U on X (i.e., remove the role of X) and then regress the residuals $r_Y = Y - \hat{Y}$ on $r_T = T - \hat{T}$ (which by FWL, is algebraically equivalent to what we would get for $\hat{\beta}$ in the full regression without U) with omitted variable $r_U = U - \hat{U}$. Using the same result as above, the omitted variable bias for just δ is then, again assuming univariate X, T, U,

$$E(\hat{\beta}|X,T,U) - \beta = \frac{\widehat{Cov}(r_T, r_U)}{\widehat{Var}(r_T)}\delta$$

Note that because $\frac{1}{n-1}$ cancels in the numerator and denominator and because the sample mean of the residuals is always 0,

$$\frac{\widehat{Cov}(r_T, r_U)}{\widehat{Var}(r_T)} = \frac{\sum_{i=1}^n (T_i - \hat{T}_i)(U_i - \hat{U}_i)}{\sum_{i=1}^n (T_i - \hat{T}_i)^2}$$

An aside: This kind of formula can be a powerful tool. Imagine, for example, that instead of fitting $Y = X\gamma + T\beta + \eta$, we fit $Y = X\gamma + T\beta + W\alpha + \eta$ where W is some imperfectly correlated proxy for U. We could again apply the FWL theorem, regressing Y, T, U on X and W and get the same kind of formula as above. Then, the more correlated W is with U, the smaller r_U will tend to be with less 'leftover variation' and hence the smaller the above covariance will be. Intuitively, if we can include a proxy for our omitted variable, we can reduce our omitted variable bias! In the extreme of course, if W and U are perfectly correlated, we do not really have an ommitted variable at all.

Notation Note: In Professor Imai's lecture slides, he uses the notation $\mathbb{V}(T^{\perp X})$ and $\text{Cov}(\mathbb{V}(T^{\perp X}), \mathbb{V}(U^{\perp X})$ to refer to the above quantities, where the $\perp$ subscript represents regressing on X and taking the residual (the orthogonal left-over part).[74]

Distinction: exogeneity vs linearity assumption Note that exogeneity is distinct from the linearity assumption. If the true model is $Y_i = e^{\alpha X} + \eta_i$ (non-linear), it could still be true that $E(\eta_i|X) = 0$. Of course, if we posed $Y_i = X_i\beta + \epsilon_i$, our model would be mis-specified and the error term $\epsilon_i = Y_i - X_i\beta$ would have expectation $E(\epsilon_i|X) = e^{\alpha X_i} - X_i\beta$, which need not be 0, but the primary problem here would be mis-specification. We could fit the model $Y_i = e^{\alpha X} + \eta_i$ and say exogeneity holds.

C.3 Heteroskedasticity (A3)

If Y_i's are still **independent** but with possibly different (heteroskedastic) variances $\sigma_1^2, ..., \sigma_n^2$, then letting $\Sigma = \text{diag}(\sigma_i^2)$, we have

$$Var(\hat{\beta}|X) = (X^TX)^{-1}X^T\Sigma X(X^TX)^{-1}$$

which does not simplify to the simpler form above. Two heteroskedasticity-robust variance estimators that come up in the course are:

1. EHW: $\hat{V}_{EHW}(\hat{\beta}|X) = (X^TX)^{-1}X^T\text{diag}(r_i^2)X(X^TX)^{-1}$ This estimator is consistent.[75]
2. HC2: $\hat{V}_{HC2}(\hat{\beta}|X) = (X^TX)^{-1}(X^T\text{diag}\left(\frac{r_i^2}{1-h_{ii}}\right)X)(X^TX)^{-1}$ where h_{ii} are diagonal elements of H (you may have encountered these as the **leverage**) and $h_{ii} = x_i^T(X^TX)^{-1}x_i$. This estimator is unbiased.

In the course, you will be able to call `R` packages that implement these, but it is good to have some sense of why they are needed. Intuitively, both are using the squared residual as an estimate of the variance of Y_i. The HC2 estimator recognizes that the variance of residuals has a relationship to leverage (above: $Var(R) = \sigma^2(I - H)$) and therefore uses leverage as an adjustment.

[74] There is a slight ambiguity where sometimes, these quantities refer to sample variances and sometimes to theoretical variances, so I add a $\hat{V}(T^{\perp X})$ for the sample quantity in my section notes.

[75] Note: the notation $\text{diag}(a_i)$ means a square matrix with $a_1, ..., a_n$ on the diagonal and 0's otherwise.

References

Abadie, A. (2005, 01). Semiparametric Difference-in-Differences Estimators. *The Review of Economic Studies 72*(1), 1–19.

Abadie, A., A. Diamond, and J. Hainmueller (2010). 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.

Abadie, A. and J. Gardeazabal (2003). The economic costs of conflict: A case study of the basque country. *The American Economic Review 93*(1), 113–132.

Abadie, A. and G. W. Imbens (2011). Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics 29*(1), 1–11.

Abadie, A. and J. L'Hour (2021). A penalized synthetic control estimator for disaggregated data. *Journal of the American Statistical Association 116*(536), 1817–1834.

Acharya, A., M. Blackwell, and M. Sen (2016). Explaining causal findings without bias: Detecting and assessing direct effects. *American Political Science Review 110*(3), 512–529.

Agresti, A. (2015). *Foundations of linear and generalized linear models*. Wiley series in probability and statistics. Hoboken, New Jersey: John Wiley & Sons Inc.

Balke, A. and J. Pearl (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association 92*(439), 1171–1176.

Basse, G. and I. Bojinov (2020). A general theory of identification.

Bell, R. and D. McCaffrey (2002). Bias reduction in standard errors for linear regression with multi-stage samples. *Survey Methodology 28*, 169–182.

Ben-Michael, E., A. Feller, and J. Rothstein (2019). Synthetic controls with staggered adoption.

Card, D. and A. B. Krueger (1993, October). Minimum wages and employment: A case study of the fast food industry in new jersey and pennsylvania. Working Paper 4509, National Bureau of Economic Research.

Cinelli, C. and C. Hazlett (2019, 12). Making Sense of Sensitivity: Extending Omitted Variable Bias. *Journal of the Royal Statistical Society Series B: Statistical Methodology 82*(1), 39–67.

Cornfield, J., W. Haenszel, E. C. Hammond, A. M. Lilienfeld, M. B. Shimkin, and E. L. Wynder (1959, 01). Smoking and Lung Cancer: Recent Evidence and a Discussion of Some Questions. *JNCI: Journal of the National Cancer Institute 22*(1), 173–203.

Cunningham, S. (2021). *Causal Inference: The Mixtape*. Yale University Press.

Ding, P. and T. J. Vanderweele (2014, 08). Generalized Cornfield conditions for the risk difference. *Biometrika 101*(4), 971–977.

Doudchenko, N. and G. W. Imbens (2016, October). Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. National Bureau of Economic Research.

Dunning, T. and J. Nilekani (2013). Ethnic quotas and political mobilization: Caste, parties, and distribution in indian village councils. *American Political Science Review 107*(1), 35–56.

Eli Ben-Michael, A. F. and J. Rothstein (2021). The augmented synthetic control method. *Journal of the American Statistical Association 116*(536), 1789–1803.

Fan, J., K. Imai, I. Lee, H. Liu, Y. Ning, and X. Yang (2023). Optimal covariate balancing conditions in propensity score estimation. *Journal of Business & Economic Statistics 41*(1), 97–110.

Fan Li, K. L. M. and A. M. Zaslavsky (2018). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association 113*(521), 390–400.

Feller, A., F. Mealli, and L. Miratrix (2017). Principal score methods: Assumptions, extensions, and practical considerations. *Journal of Educational and Behavioral Statistics 42*(6), 726–758.

Gelman, A. and G. Imbens (2019). Why high-order polynomials should not be used in regression discontinuity designs. *Journal of Business & Economic Statistics 37*(3), 447–456.

Gerard, F., M. Rokkanen, and C. Rothe (2020). Bounds on treatment effects in regression discontinuity designs with a manipulated running variable. *Quantitative Economics 11*(3), 839–870.

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis 20*(1), 25–46.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica 47*(1), 153–161.

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

Imai, K. and I. S. Kim (2019). When should we use unit fixed effects regression models for causal inference with longitudinal data? *American Journal of Political Science 63*(2), 467–490.

Imai, K., I. S. Kim, and E. H. Wang (2023). Matching methods for causal inference with time-series cross-sectional data. *American Journal of Political Science 67*(3), 587–605.

Imbens, G. W. (2020). Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. *Journal of Economic Literature 58*(4), pp. 1129–1179.

Imbens, G. W. and D. B. Rubin (2015). *Causal inference for statistics, social, and biomedical sciences : an introduction.* New York: Cambridge University Press.

King, G. and R. Nielsen (2019). Why propensity scores should not be used for matching. *Political Analysis 27*(4), 435–454.

Liang, K.-Y. and S. L. Zeger (1986). Longitudinal data analysis using generalized linear models. *Biometrika 73*(1), 13–22.

Manski, C. F. (1990). Nonparametric bounds on treatment effects. *The American Economic Review 80*(2), 319–323.

Niknam, B. A. and J. R. Zubizarreta (2022, 01). Using Cardinality Matching to Design Balanced and Representative Samples for Observational Studies. *JAMA 327*(2), 173–174.

Pang, X., L. Liu, and Y. Xu (2022). A bayesian alternative to synthetic control for comparative case studies. *Political Analysis 30*(2), 269–288.

Rawlings, J. O. (1998). *Applied regression analysis : a research tool* (Second edition. ed.). Springer texts in statistics. New York: Springer.

Robins, J. M. (1994). Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics - Theory and Methods 23*(8), 2379–2412.

Rosenbaum, P. R. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association 84*(408), 1024–1032.

Sant'Anna, P. H. and J. Zhao (2020). Doubly robust difference-in-differences estimators. *Journal of Econometrics 219*(1), 101–122.

Shen, D., P. Ding, J. Sekhon, and B. Yu (2022). Same root different leaves: Time series and cross-sectional methods in panel data.

van der Vaart, A. W. (1998). *Asymptotic statistics.* Cambridge series on statistical and probabilistic mathematics ; 3. Cambridge: Cambridge University Press.

Victor Chernozhukov, K. W. and Y. Zhu (2021). An exact and robust conformal inference method for counterfactual and synthetic controls. *Journal of the American Statistical Association 116*(536), 1849–1864.

Visconti, G. and J. R. Zubizarreta (2018). Handling limited overlap in observational studies with cardinality matching. *Observational Studies 4*(1), 217–249. Submitted 11/17; Published 7/18.

Wang, Y. and R. D. Shah (2020). Debiased inverse propensity score weighting for estimation of average treatment effects with high-dimensional confounders.

Wellek, S. (2010). *Testing Statistical Hypotheses of Equivalence and Noninferiority* (2 ed.). Chapman and Hall/CRC.

Wood, G., T. R. Tyler, and A. V. Papachristos (2020). Procedural justice training reduces police use of force and complaints against officers. *Proceedings of the National Academy of Sciences 117*(18), 9815–9821.

Zubizarreta, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association 110*(511), 910–922.