Lecture 2 -Research Designs and Model vs. Design-Based Causal Inference

Paul Goldsmith-Pinkham

January 18, 2024

There are three goals for this set of notes:

1. Discuss the value of randomized interventions, and more generally identifying settings where interventions are "as-if" randomly assigned. In doing so, we'll touch on the historical and (somewhat) current views on this.
2. Define a "research design."
3. Give an introduction to design-based vs. model-based identification and causal inference.

Randomization

Randomization is a powerful tool. Being able to truly randomize an intervention allows the researcher to assume (by definition) that the potential outcomes for units are independent, satisfying the first assumption in strong ignorability.

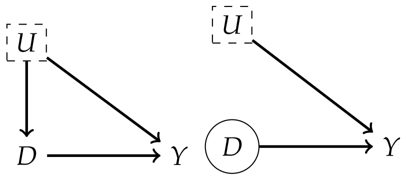


Figure 1: D's effect on Y is confounded by U, but a randomized intervention of D breaks any back-door connection

In a DAG, randomization ensures that any backdoor path to D is broken, since the randomization was the only cause of the intervention. This allows identification of the total effect on Y.[1]

If the use of randomization is so powerful, why don't we always use it? There are a few reasons:

1. People may not want to be randomized into different treatments. They value their choices, and it may be impractical to randomize their decisions even if there is a clear benefit to doing so. A firm, for example, may not want to randomize their policies (although they want to in some cases, as in settings with A/B testing).
2. It may be unethical to randomize. For example, if there is a clear benefit to a treatment, it may be unethical to withhold that treatment from individuals by placing them in the control.[2]

[1] Randomization does not necessarily identify the direct effect of D on Y. For example, if D affects multiple outcomes, X and Y, and then X affects Y as well, it's possible that agents may reoptimize their X, thereby offsetting (or increasing) the direct effect of D on Y.

[2] The concept of "equipoise" is often used to describe the ethical considerations of randomization in the medical literature [Freedman, 1987]: "if there is genuine uncertainty within the expert medical community — not necessarily on the part of the individual investigator — about the preferred treatment."

3. It may be impossible to randomize. For example, if we are interested in the effect of a policy change, it may be impossible to randomize the policy change across different regions or states.

The credibility revolution – then and now

While randomization is often viewed as the gold standard for policy evaluation, this was not always the case. In fact, the use of randomized experiments in economics is relatively new. Indeed, while there was the occasional agricultural economics application that had true randomization, most econometric modeling estimating causal effects and structural parameters was based on arguments about models and controls. This led to substantial skepticism in the broader community by the end of the 1970s. This can be seen in discussion about econometric estimates in Leamer [1983], "Let's take the con out of econometrics":[3]

> After three decades of churning out estimates, the econometrics club finds itself under critical scrutiny and faces incredulity as never before. Fischer Black writes of "The Trouble with Econometric Models." David Hendry queries "Econometrics: Alchemy or Science?" John W. Pratt and Robert Schlaifer question our understanding of "The Nature and Discovery of Structure." And Christopher Sims suggests blending "Macroeconomics and Reality.

Quite explicitly, Black [1982] says: "The trouble with econometric models is that they present correlations disguised as causal relations. The more obvious confusions between correlation and causation can often be avoided, but there are many subtle ways to confuse the two; in particular, the language of econometrics encourages this confusion."

The state of applied research is summarized (in a somewhat extreme way) by Leamer as:

> Econometricians would like to project the image of agricultural experimenters who divide a farm into a set of smaller plots of land and who select randomly the level of fertilizer to be used on each plot. If some plots are assigned a certain amount of fertilizer while others are assigned none, then the difference between the mean yield of the fertilized plots and the mean yield of the unfertilized plots is a measure of the effect of fertilizer on agricultural yields. The econometrician's humble job is only to determine if that difference is large enough to suggest a real effect of fertilizer, or is so small that it is more likely due to random variation.
>
> This image of the applied econometrician's art is grossly misleading. I would like to suggest a more accurate one. **The applied econometrician is like a farmer who notices that the yield is somewhat higher under trees where birds roost, and he uses this as evidence that bird**

[3] This Leamer [1983] article is really worth reading in full.

> **droppings increase yields.** However, when he presents this finding at the annual meeting of the American Ecological Association, another farmer in the audience objects that he used the same data but came up with the conclusion that moderate amounts of shade increase yields. A bright chap in the back of the room then observes that these two hypotheses are indistinguishable, given the available data. He mentions the phrase "identification problem," which, though no one knows quite what he means, is said with such authority that it is totally convincing.[4]

Finally, Leamer argues that the reason randomization is so helpful is that it removes the need to arbitrarily try many specifications to check for robustness to other confounding causes:

> The truly sharp distinction between inference from experimental and inference from nonexperimental data is that experimental inference sensibly admits a conventional horizon in a critical dimension, namely the choice of explanatory variables. If fertilizer is randomly assigned to plots of land, it is conventional to restrict attention to the relationship between yield and fertilizer, and to proceed as if the model were perfectly specified... In contrast, it would be foolhardy to adopt such a limited horizon with nonexperimental data. **But if you decide to include light level in your horizon, then why not rainfall; and if rainfall, then why not temperature; and if temperature, then why not soil depth, and if soil depth, then why not the soil grade; ad infinitum.** Though this list is never ending, it can be made so long that a nonexperimental researcher can feel as comfortable as an experimental researcher that the risk of having his findings upset by an extension of the horizon is very low. The exact point where the list is terminated must be whimsical, but the inferences can be expected not to be sensitive to the termination point if the horizon is wide enough.

If we fast-forward 25 years, Angrist and Pischke [2010] have now declared victory: "The credibility revolution in empirical economics: How better research design is taking the con out of econometrics":

> Empirical microeconomics has experienced a credibility revolution, with a consequent increase in policy relevance and scientific impact. Sensitivity analysis played a role in this, but as we see it, the primary engine driving improvement has been a focus on the quality of empirical research designs... The advantages of a good research design are perhaps most easily apparent in research using random assignment, which not coincidentally includes some of the most influential microeconometric studies to appear in recent years.

As evidenced by both the title and the quote above, **research design** is declared the victor. But what is a research design? And why is randomization its champion?

[4] It continues: "The meeting reconvenes in the halls and in the bars, with heated discussion whether this is the kind of work that merits promotion from Associate to Full Farmer; the Luminists strongly opposed to promotion and the Aviophiles equally strong in favor."

Example 1
Some famous examples of programs that used randomization (including those that they cite) include:

- *PROGRESA, a conditional cash transfer program in Mexico (see Parker and Todd [2017] for a review)*
- *Moving to Opportunity (MTO), a program that randomly selected low income families to receive housing vouchers (see Katz et al. [2001] for a discussion on the program)*
- *National Supported Work (NSW) demonstration, a federal job training program that was randomized amongst applicants (LaLonde [1986] is the canonical paper whose work with this program is what sparked the "credibility revolution" – we will discuss this next class)*
- *Oregon health insruance experiment, where the state of Oregon randomized its Medicaid program for low-income, uninsured adults. See Baicker et al. [2013] for a discussion.*
- *Tennessee STAR class size experiment, which randomized students into classrooms of different sizes. See Krueger [1999] for a discussion.*
- *H&R Block FAFSA was field experiment where individuals recieving tax preperation at H&R Block were randomized into a procedure to get help on the Free Application for Federal Student Aid (FAFSA). See Bettinger et al. [2012] for a discussion.*

What is a research design?

A goal of this class, and your empirical research going forward, is to have a precise research design for your empirical analyses. This is a term that is used frequently, but not always clearly defined. In fact, in the Angrist and Pischke [2010] paper, the term is never defined explicitly, despite being mentioned 69 times. I have seen it defined explicitly only a few times, and rarely in economics.[5]

I will provide you a definition, with the understanding that this is not the only definition, and that there are many different ways to think about research design. Much of the value in thinking about a research design is about being explicit about the assumptions that you are making, and how you are using the data to answer your question.

A research design is a statistical and/or economic statement of

[5] One very nice text in political science that does define it is Blair et al. [2023]. They define a research design as "a procedure for generating answers to questions."

how to estimate a causal relationship between two variables of interest: how X causes Y. Since we know that causal effects require the estimation of an (unobservable) counterfactual, this statement describes the assumptions necessary to impute the counterfactual. Why is this valuable?

First, it forces you to articulate *what* the counterfactual is. This may seem obvious, but often you may find researchers estimating a linear equation and presenting estimates without clearly thinking about their counterfactual statement. For example, when you estimate the effect of a policy change, what is the counterfactual? Is it the state of the world where the policy never occured? Or is it one where the policy was introduced later? Or when estimating the effect of an informational event (such as the effect of monetary policy), is the counterfactual where the event never occurred? Or is it where the event occurred as previously expected?

Second, it forces you to articulate *how* you are going to estimate the counterfactual, and what assumptions are necessary. This is, of course, what we will spend the rest of the semester building tools to do. But at a very high-level, a research design can be split into two types of approaches: model-based and design-based. Model-based approaches will involve assumptions about modeling the expectation (or other functional) of the counterfactual, specifically dealing with any possible confounding variables. Design-based approaches will involve assumptions about the treatment assignment mechanism, without making formal assumptions about the model of the potential outcomes.

Comment 1

- *Model-based: the estimand is identified using assumptions on the modeling of the potential outcomes conditional on treatment and additional variables (e.g. parallel trends). Examples of approaches that can fall under this category include difference-in-differences (including non-random staggerred diff-in-diff), regression discontinuity, synthetic control (and synthetic diff-in-diff), and instrumental variable approaches that use "included" instruments.*

- *Design-based: the estimand is identified using assumptions on the treatment variable, conditional on the potential outcomes and additional variables. Examples include randomized control trials, instrumental variable approaches that use "excluded" instruments, difference-in-difference with* random *staggered timing, and propensity score matching.*

See Lihua Lei's very nice twitter thread for a small history on why these terms acquired their labels.

To give a concrete example of how these assumptions may differ, we can use the example from Robins et al. [1992]. Consider the question of how smoking affects peoples' ability to breath, as measured by "forced expiratory volume in one second" (FEV1). This is often used as a measure of lung function. Now we want to know what the effect of a person being a smoker (D_i) is on the individuals' FEV1 (Y_i). The two approaches (model and design) highlight the different ways you might consider estimating the effect. One approach would be to think hard about ways that shift around an individuals' propensity to be a smoker in as-if random ways – this would be a *design* approach, since it is focused on the treatment assignment mechanism. Another approach might be to compare individuals over time in places where cigarette smoking was legal earlier vs. later – this would be a *model* approach, since it is focused on the modeling of the potential outcomes by using the individuals in the state with later smoking as a control for the earlier group.[6]

As it turns out, not only do these approaches matter for clarity of thought, they matter for robustness of estimation (design-based inference will be robust to model specification), weighting of estimands (model-based approaches will be more sensitive to negative weights), and the ability to generalize to other settings (model-based approaches are often more easily generalized, conditional on the model being correct). Moreover, one approach, with the same research data and causal question, may be much more statistically

[6] One could take the *same* data and use it for both approaches. If you were willing to assume that the states chose to make the cigarettes legal late vs. early randomly, then this would be a design-based approach, since it would influence the treatment assignment mechanism.

precise than another. We will continue to explore and describe these two approaches throughout the semester.

Randomization and design-based inference

Returning to randomization, we can see that randomized interventions are a form of design-based causal inference. Konwledge of the treatment assignment mechanism gives a very powerful tool for thinking about the counterfactual. In fact, it is so powerful that it is the benchmark for other approaches in design-based inference. That is, a randomized intervention with knowledge of the treatment assignment mechanism is the "gold standard." In future cases, we will need to make assumptions about the treatment assignment mechanism and defend them. For now, we will provide the notation and estimators for the randomized case, and next class we will discuss more general approaches.

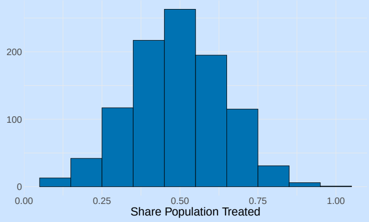
As before, there is a finite population of n individuals indexed by i. For each i, we have triplets $(Y_i(0), Y_i(1), D_i)$, where $D_i \in \{0,1\}$ is the treatment status and $(Y_i(0), Y_i(1))$ denotes the potential outcomes. We observe (Y_i, D_i), and define the vector version of these as $\mathbf{Y}$ and $\mathbf{D}$. There are many things we could want to know abotu the relationship between D_i and $\tau_i = Y_i(1) - Y_i(0)$, but for today, we will focus on $\bar{\tau} = n^{-1} \sum_{i=1}^{n} \tau_i$.[7]

Design-based inference considers the set of potential ways that $\mathbf{D}$ could be randomized to the population. We will assume that $\mathbf{Y}_1$ and $\mathbf{Y}_0$ are *fixed* – it is only the random variation in $\mathbf{D}$ that creates uncertainty. Formally, let Ω denote that space of possible values that $\mathbf{D}$ can take. It is defined by the type of randomized experiment one runs.

[7] One could, for example, study the median treatment effect, or other features of the distribution. This is more complex, as we will see in future lectures.

Example 2

If we do a purely randomized individualized trial, where each individual has a fair coin flipped on whether they are treatment or control, then $\Omega = \{0,1\}^n$. *But then the variation in number treated and control can vary quite a lot for small samples!*



Other ways to consider randomly assigning individuals include:

- *Random draws from an urn (to ensure an exact number treated)*
- *Clustering individuals on characteristics (or location)*

Given our sample space and knowledge of the randomziation, we know the exact probability distribution over Ω, and hence $\mathbf{D}$.

Example 3

Consider a sample of 10 units, with 5 treated and 5 control. We know that there are only $\binom{10}{5} = 252$ *potential combinations (each equally likely). We observe one set of them in Table 1. Note that one set of the entries (in blue) are fundamentally unobservable due to the treatment status.*

D_i	$Y_i(1)$	$Y_i(0)$	Y_i	τ_i
1	11.9	6.6	11.9	5.3
1	10.0	8.5	10.0	1.5
1	9.7	9.4	9.7	0.3
1	9.5	7.0	9.5	2.5
1	11.4	7.4	11.4	4.0
0	9.6	7.6	7.6	2.0
0	9.1	7.1	7.1	2.0
0	10.4	7.7	7.7	2.7
0	10.4	8.0	8.0	2.4
0	12.4	7.8	7.8	4.6

Table 1: Example of a randomization over $n = 10$ units. The highlighted entries are unobservable due to the fundamental problem of causal inference.

Now, we need an estimator for $\bar{\tau} = n^{-1}\sum_{i=1}^{n}\tau_i$. We already know under random assignment that $E(Y_i|D_i = 1) - E(Y_i|D_i = 0)$ identifies $E(\tau_i)$. Then, the empirical analog is quite easy (with n_1 equal to the number of treated, n_0 number control, and $n_0 + n_1 = n$):

$$\hat{\bar{\tau}}(\mathbf{D},\mathbf{Y}) = \frac{\mathbf{D}'\mathbf{Y}}{\sum_i D_i} - \frac{(\mathbf{1}-\mathbf{D})'\mathbf{Y}}{\sum_i(1-D_i)} \tag{1}$$

$$= n_1^{-1}\sum_i Y_i D_i - n_0^{-1}\sum_i Y_i(1-D_i) \tag{2}$$

Note that this expectation operator is well-defined from the objects we already know. Since only D is random, and we know its marginal distribution over the sample we can show that this estimator is unbiased. This particular estimator is unbiased when the *design*, or the randomization across Ω, is special: it has complete random assignment across of the units across the treatment. We assume that n_1/n units are randomly assigned (in our example in Table 1, $5/10$.).[8]

Under this design,the probability of a unit receiving treatment

[8] If the probabiltiies vary across the sample space (due to covariates, say, or a more unusual sampling scheme), we need to add weights. This is known as Horovitz-Thompson weighting, and we will return to this.

given a draw $\mathbf{D}$ is $\pi = n_1/n$. Note that this implies that there are always n_1 units treated, and we are randomly allocating the treatments within the n units. Then, note that $E(\pi_1^{-1} D_i) = 1$.[9] With this,

$$\begin{aligned}
E(\hat{\hat{\tau}}(\mathbf{D}, \mathbf{Y})) &= E\left(\frac{\mathbf{D}'\mathbf{Y}}{\sum_i D_i} - \frac{(\mathbf{1}-\mathbf{D})'\mathbf{Y}}{\sum_i (1-D_i)}\right) \\
&= n^{-1} E\left(\sum_i \pi_1^{-1} Y_i D_i - \sum_i (1-\pi_1)^{-1} Y_i (1-D_i)\right) \\
&= n^{-1} E\left(\sum_i \pi_1^{-1} Y_i(1) D_i - \sum_i (1-\pi_1)^{-1} Y_i(0)(1-D_i)\right) \\
&= n^{-1} \sum_i Y_i(1) E\left(\pi_1^{-1} D_i\right) - n^{-1} \sum_i Y_i(0) E\left((1-\pi_1)^{-1}(1-D_i)\right) \\
&= n^{-1} \sum_i Y_i(1) - Y_i(0) = n^{-1} \sum_i \tau_i.
\end{aligned}$$

Hence, this estimator is unbiased for the ATE.

We can also study the variance properties of the estimator. Thanks to Splawa-Neyman et al. [1990], we know that the variance of $\hat{\hat{\tau}}$ is given by:

$$\sigma^2_{\hat{\hat{\tau}}} = \frac{1}{n-1}\left(\frac{n_1 \sigma_0^2}{n_0} + \frac{n_0 \sigma_1^2}{n_1} + 2\sigma_{0,1}\right) \tag{3}$$

where $\sigma_0^2, \sigma_1^2, \sigma_{0,1}$ are the variance of the potential control, treatment, and the covariance between the two. Note that these variances are of the potential outcomes. Some nice intuition can come from looking at this. First, we see that the variance of the estimator increases when either the treated or control variance increases. This makes sense – it is harder to distinguish treatment and control when there is a large dispersion for either group. Second, the overall variance is increases (holding fixed the specific variances) as you increase the share of treated units. This makes sense because you have less information about the control for the treatment. Finally, the covariance of the potential outcomes matters for the overall variance – if the units have negative covariance, that will help in estimating the treatment effect because a large shock to the control potential outcome will be offset by a large shock in the other direction for the treatment.

Since we do not know $\sigma_{0,1}$, we need to bound this estimand with a conservative estimator:

$$\hat{\sigma}^2_{\hat{\hat{\tau}}} = \frac{n}{n-1}\left(\frac{\hat{\sigma}_0^2}{n_0} + \frac{\hat{\sigma}_1^2}{n_1}\right). \tag{4}$$

This estimator is knowable from the data, if the treatment is randomly assigned.

[9] Recall that $E(D_i) = Pr(D_i = 1) = n^{-1} n_1$.

Example 3 (continued)
We can now construct our estimator and the variance of this estimator:

$$\hat{\hat{\tau}} = 5^{-1}\sum_i Y_i D_i - 5^{-1}\sum_i Y_i(1-D_i) = 2.86$$

and

$$\hat{\sigma}_0^2 = 5^{-1}\sum_i (Y_i - \hat{\hat{Y}}_0)^2(1-D_i) = 0.932$$

$$\hat{\sigma}_1^2 = 5^{-1}\sum_i (Y_i - \hat{\hat{Y}}_1)D_i = 0.0904$$

$$\hat{\sigma}_{\hat{\tau}}^2 = \frac{10}{9}\left(\frac{0.932}{5} + \frac{0.0904}{5}\right) = 0.2$$

Hence, our standard error is $\sqrt{0.2} = 0.45$.

Comment 2
It is interesting to note that this variance estimator is nearly identical to the case with the standard robust estimator from a more traditional linear equation:

$$Y_i = \alpha + \beta D_i + \epsilon_i.$$

See Equation 2 from Imbens and Kolesar [2016] to compare.

Thinking about inference

We could use this variance estimator to thinking about constructing *confidence intervals* now. Often, this is done by inverting a hypothesis test. For example, we could test the null hypothesis that $E(\tau_i) = 0$ – the average treatment effect in the sample is zero. We will revisit this in further detail in our linear regression classes, and you have likely seen quite a bit of this in your previous classes. Thinking about testing in the design-based setting will be no different – the only change is that the uncertainty is driven by the random assignment of the treatment, rather than uncertainty in the outcome (e.g. usually the errors in the model). It is not always easy to figure out *what* the variance of an estimator is that has a non-standard design. We will discuss simple cases where probabilities are done in a straightforward way, but often experiments are run in ways that create unusual dependence across units.[10]

One very powerful tool that can avoid estimation of standard errors is to use randomization inference instead. One example where we can use this is in testing the strong null hypothesis that $\tau_i = 0$ for

[10] See Imbens and Rubin [2015] for a general discusion and Chang [2023] for a discussion on complex experiements.

all i. That is, the treatment has zero effect. This is a very strong null hypothesis – it is stronger than the null hypothesis that $\bar{\tau} = 0$.

Given our data and under the null of $\tau_i = 0$, we can calculate the full distribution of potential observed statistics we would see, as we vary D. We do so by imputing our missing values under the null hypothesis, and calculating the estimator if we randomly permuted the treatment labels. Since we are asserting the known missing values, we can reconstruct the full distribution in Figure 2. We can then calculate the probability of seeing a value as extreme as our observed value. This is known as a *p-value*. If this probability is small, we reject the null hypothesis that $\tau_i = 0$ for all i.

D_i	$Y_i(1)$	$Y_i(0)$	Y_i
1	11.9	11.9	11.9
1	10	10	10
1	9.7	9.7	9.7
1	9.5	9.5	9.5
1	11.4	11.4	11.4
0	7.6	7.6	7.6
0	7.1	7.1	7.1
0	7.7	7.7	7.7
0	8	8	8
0	7.8	7.8	7.8

Table 2: Imputed values under the null hypothesis of $\tau_i = 0$ for all i.

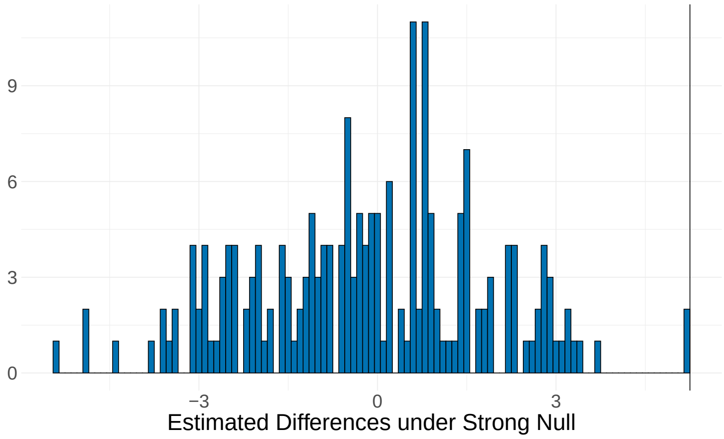


Figure 2: Distribution of $\hat{\bar{\tau}}$ under the null hypothesis of $\tau_i = 0$ for all i under all permutations. Vertical line denotes the observed estimate in the data.

Comment 3

We have only discussed a very simple estimator which assumes complete randomization. The generalized estimator that allows for more complex randomization schemes is known as the Horvitz-Thompson estimator from Horvitz-Thompson (1952) (see Aronow and Middleton (2013) for a useful discussion):

$$\hat{\bar{\tau}}_{HT} = n^{-1}\left[\sum_i \frac{1}{\pi_{1i}} Y_i D_i - \frac{1}{\pi_{0i}} Y_i (1 - D_i)\right], \tag{5}$$

where $\pi_{i1} = Pr(D_i = 1)$, and $\pi_{0i} = Pr(D_i = 0)$. This estimator is unbiased even in settings where we don't have equal weighting across the sampling space. This is reweighting using the propensity score! We will discuss this next class.

Credibility revolution and internal vs. external validty

The focus on randomization and credible design has had an extremely powerful impact of the believability of estimates. However, there was (and sometimes is) a view that the emphasis in these approaches focuses too much on solving problems of *internal validity* (i.e. the ability to identify the causal effect *in the sample*) and not enough on *external validity* (i.e. the ability to generalize to other settings).

This debate around internal vs. external validity erupted at the end of the 2000s, especially focused in development economics. Papers in this space include:

- "Instruments, Randomization, and Learning about Development" Deaton (2010)
- "Comparing IV with structural models: What simple IV can and cannot identify", Heckman and Urzua (2009)
- "Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)" Imbens (2010)
- "Building Bridges between Structural and Program Evaluation Approaches to Evaluating Policy" Heckman (2010)

Much of this is tied to instrumental variables, which we'll revisit later. To give you a flavor of the issue as flagged in development, here is Angus Deaton in 2010 [Deaton, 2010] describing issues with randomized experiments in development:

> Under ideal circumstances, randomized eval uations of projects are useful for obtaining a convincing estimate of the average effect of a program or project. The price for this success is a focus that is too narrow and too local to tell us "what works" in development, to design policy, or to advance scientific knowledge about development processes. Project evaluations, whether using random ized controlled trials or nonexperimental methods, are unlikely to disclose the secrets of development nor, unless they are guided by theory that is itself open to revision, are they likely to be the basis for a cumulative research program that might lead to a better understanding of development.

As another example, here is a table from Heckman [2010] that compares the assumptions needed for potential outcomes vs. structural work (a dichotomy which I think is somewhat vacuous), and emphasizing the external validity problem in potential outcomes work:[11]

Many of the complaints by the anti-randomistas devolve into three types: first, the analyses are done incorrectly (e.g. bad IVs). I think full-throated defenders of experiments would agree that badly done research should be rejected regardless. More importantly, the transparency of the research design should make this easier. Second, that research does not generalize to other populations. For example, Progressa is a big success, but knowing that conditional cash transfers

[11] There are a number of statements in this table that are not correct regarding potential outcomes. For example, the statement that social interactions is assumed away is not correct. See Lecture 4 for more discussion.

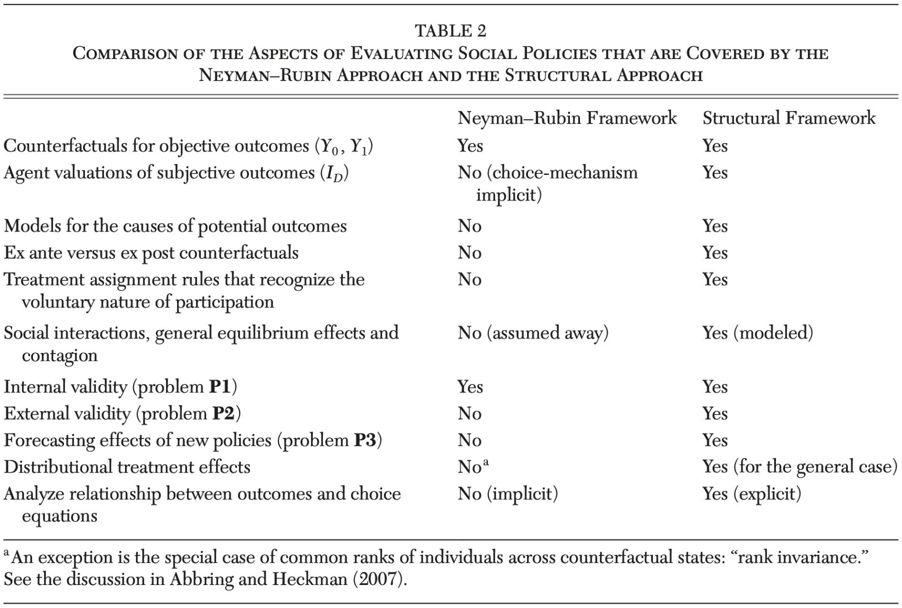


TABLE 2
COMPARISON OF THE ASPECTS OF EVALUATING SOCIAL POLICIES THAT ARE COVERED BY THE NEYMAN–RUBIN APPROACH AND THE STRUCTURAL APPROACH

	Neyman–Rubin Framework	Structural Framework
Counterfactuals for objective outcomes (Y_0, Y_1)	Yes	Yes
Agent valuations of subjective outcomes (I_D)	No (choice-mechanism implicit)	Yes
Models for the causes of potential outcomes	No	Yes
Ex ante versus ex post counterfactuals	No	Yes
Treatment assignment rules that recognize the voluntary nature of participation	No	Yes
Social interactions, general equilibrium effects and contagion	No (assumed away)	Yes (modeled)
Internal validity (problem **P1**)	Yes	Yes
External validity (problem **P2**)	No	Yes
Forecasting effects of new policies (problem **P3**)	No	Yes
Distributional treatment effects	No[a]	Yes (for the general case)
Analyze relationship between outcomes and choice equations	No (implicit)	Yes (explicit)

[a] An exception is the special case of common ranks of individuals across counterfactual states: "rank invariance." See the discussion in Abbring and Heckman (2007).

work in this one setting may not necessarily inform our ability to roll it out in places that are very different. Third, that there is a rhetorical overreliance on RCTs as the gold standard, and that post-hoc analyses (without a pre-analysis plan) defeat the underlying value of an RCT anyway.[12] More generally, there is a concern that focusing on clever RCTs and IVs causes an overfocus on irrelevant or unimportant questions. A briefcase full of results that are not economically useful.[13]

It is useful to consider these concerns in the context of the discussion at the beginning of this lecture. Much of this concern about how to do empirical work does not provide much of a counterfactual. Historical evidence suggests that empirical work was simply not credible prior to this move. Additionally, it seems like the concerns about empirics being too separated from models are overstated. Perhaps in part in response to these critiques, many empirical papers with causal parameters are tightly linked to theretical work. For those that are not, the results eventually inform many theoretical papers. A push to open data has actually made it easier for researchers to follow-up and study these issues.

The key way in which "better research design is taking the con out of econometrics" is by making the assumptions in empirical work *explicit*. This can be using a randomized intervention, or some other design-based approach, or it can be done using a model-based approach. Then, researchers can evaluate clearly the credibility of the

[12] It is unclear why an experiment is worse than a non-experiment in this regard, but this is a concern Deaton flags.

[13] This is still a complaint one can hear today!

assumptions, and the robustness of the results to these assumptions. *The inclusion of an economic model does not grant an empirical researcher to omit a research design from their empirics.* Many researchers may propose a model, and then demonstrate that their model is consistent with observational data. This is not a research design, which requires an additional argument for how the empirical approach can be used to identify the causal estimand of interest.

References

Joshua D Angrist and Jörn-Steffen Pischke. The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of economic perspectives*, 24(2):3–30, 2010.

Katherine Baicker, Sarah L Taubman, Heidi L Allen, Mira Bernstein, Jonathan H Gruber, Joseph P Newhouse, Eric C Schneider, Bill J Wright, Alan M Zaslavsky, and Amy N Finkelstein. The oregon experiment—effects of medicaid on clinical outcomes. *New England Journal of Medicine*, 368(18):1713–1722, 2013.

Eric P Bettinger, Bridget Terry Long, Philip Oreopoulos, and Lisa Sanbonmatsu. The role of application assistance and information in college decisions: Results from the h&r block fafsa experiment. *The Quarterly Journal of Economics*, 127(3):1205–1242, 2012.

Fischer Black. The trouble with econometric models. *Financial Analysts Journal*, 38(2):29–37, 1982.

Graeme Blair, Alexander Coppock, and Macartan Humphreys. *Research Design in the Social Sciences: Declaration, Diagnosis, and Redesign*. Princeton University Press, 2023.

Haoge Chang. Design-based estimation theory for complex experiments. *arXiv preprint arXiv:2311.06891*, 2023.

Angus Deaton. Instruments, randomization, and learning about development. *Journal of economic literature*, 48(2):424–455, 2010.

B Freedman. Equipoise and the ethics of clinical research. *The New England Journal of Medicine*, 317(3):141–145, 1987.

James J Heckman. Building bridges between structural and program evaluation approaches to evaluating policy. *Journal of Economic literature*, 48(2):356–398, 2010.

Guido W Imbens and Michal Kolesar. Robust standard errors in small samples: Some practical advice. *Review of Economics and Statistics*, 98(4):701–712, 2016.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Lawrence F Katz, Jeffrey R Kling, and Jeffrey B Liebman. Moving to opportunity in boston: Early results of a randomized mobility experiment. *The quarterly journal of economics*, 116(2):607–654, 2001.

Alan B Krueger. Experimental estimates of education production functions. *The quarterly journal of economics*, 114(2):497–532, 1999.

Robert J LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, pages 604–620, 1986.

Edward E Leamer. Let's take the con out of econometrics. *The American Economic Review*, 73(1):31–43, 1983.

Susan W Parker and Petra E Todd. Conditional cash transfers: The case of progresa/oportunidades. *Journal of Economic Literature*, 55 (3):866–915, 2017.

James M Robins, Steven D Mark, and Whitney K Newey. Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics*, pages 479–495, 1992.

Jerzy Splawa-Neyman, D. M. Dabrowska, and T. P. Speed. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5(4):465–472, 1990. ISSN 08834237. URL http://www.jstor.org/stable/2245382.

Lecture 3 - Propensity Scores

Paul Goldsmith-Pinkham

January 21, 2024

For today, propensity scores. The end goal:

- Have a framework for discussing subpopulations being treated
- A way to link to an underlying economic model

This will provide structure for us later

Propensity score weighting

Begin by recalling our definition of conditional strong ignorability:

Definition 1
We say that D_i is strongly ignorable conditional on a vector $\mathbf{X}_i$ if

1. $Y_i(0), Y_i(1) \perp D_i | \mathbf{X}_i$
2. $\exists \varepsilon > 0$ *such that* $\varepsilon < Pr(D_i = 1|\mathbf{X}_i) < 1 - \varepsilon$.

The important feature that we will engage with today is how conditional strong ignorability can depend on a potential complex and high-dimensional vector $\mathbf{X}$. When $\mathbf{X}$ is complex, it can be challenging to consider how to implement the ATE estimator as we constructed in our proof. Namely, estimating the CATE $\tau(x)$ for all x may be challenging.

What we will explore today is how the *propensity score* can be used to elide this problem. Let $\pi(\mathbf{X}_i) \equiv E(D_i|\mathbf{X}_i) = Pr(D_i|\mathbf{X}_i)$ be the probability of treatment conditional on $\mathbf{X}$. The propensity score is a scalar summary of the high-dimensional $\mathbf{X}$.[1]

A key result from Rosenbaum and Rubin [1983][2] is that if

$$Y_i(0), Y_i(1) \perp D_i | \mathbf{X}_i$$

holds, then so does $Y_i(0), Y_i(1) \perp D_i | \pi(\mathbf{X})$. The propensity score here acts as the coarsest possible "balancing score" such that the distribution of $\mathbf{X}$ is the same for both the treated and control groups. Why is this useful? It solves a high-dimensional problem – instead of exactly matching on many different values in $\mathbf{X}$, we only have to worry about a single scalar $\pi(\mathbf{X})$.

Conditioning on a single propensity score is a slightly weaker condition than conditioning on the full vector $\mathbf{X}$. However, it opens up new questions and estimation issues.

[1] A student of linear regression might notice that the propensity score is analagous to the auxiliary regression of a regression setup of Y_i on $\mathbf{X}_i$ and D_i. In essence, the propensity score captures the bias in the coefficient on D_i that would occur from omitting $\mathbf{X}$ from the main regression.

[2] Aptly named "The Central Role of the Propensity Score in Observational Studies for Causal Effects."

1. First, how do we estimate the propensity score? When $\mathbf{X}$ is discrete, we can estimate $\pi(\mathbf{X})$ non-parametrically by calculating $E(D_i|\mathbf{X}_i = x)$ for every x value, but that may ask quite a bit of the data. An alternative approach is to assume a model for $\pi(\mathbf{X})$, such as a logistic regression model. This is a parametric approach, but it can be more efficient if the model is correctly specified. A third approach is to use a parametric model but to include flexible terms for $\mathbf{X}$ to allow for non-linearities. This is a semi-parametric approach, and it can be more flexible than the fully parametric approach.

2. Second, once we have an estimated $\pi(\mathbf{X})$, how do we use it to construct the ATE or other estimands? If we directly treat it as a covariate, it becomes a bit challenging, as we discover in Example 1. Instead, we will use another beautiful result from Rosenbaum and Rubin [1983] that shows how we can use the Horvitz-Thompson estimator to construct the ATE using the propensity score.

Horvitz-Thompson Estimator

Recall our estimator for the average treatment effect from last lecture, the Horvitz-Thompson estimator:[3]

Definition 2
We observe a sample of (Y_i, X_i, D_i) triples for n observations. Let $\pi(\mathbf{X}_i) = Pr(D_i = 1|\mathbf{X}_i)$ be the ***propensity score*** *and define the* ***Horvitz-Thompson estimator*** *for the average treatment effect as:*

$$\hat{\tau}_{HT} = n^{-1} \sum_{i=1}^{n} \frac{D_i Y_i}{\pi_i(X_i)} - n^{-1} \sum_{i=1}^{n} \frac{(1-D_i)Y_i}{1-\pi_i(X_i)}$$

This estimator is the direct empirical analog to the following population result:

$$E(\tau_i) = E\left(\underbrace{\frac{Y_i D_i}{\pi(\mathbf{X})}}_{E(Y_i(1))} - \underbrace{\frac{Y_i(1-D_i)}{1-\pi(\mathbf{X})}}_{E(Y_i(0))} \right)$$

This problem takes advantage of the estimand of interest, rather than tryign to address the problem literaly by estimating every CATE and weighting up accordingly.[4]

[3] This is sometimes referred to as the inverse propensity score (or weighting) (IPW) estimator. But if you want to seem fancy you can call it the HOrvitz-Thompson estimator. Impress your statistics colleagues!

[4] Convince yourself that under discrete and few X, this collapses to what we would logically do anyway. With many X, you typically may be willing to make modeling assumptions on π for efficiency reasons.

Example 1 (Propensity score matching)
Consider the following example from Aronow and Miller [2019], with 6 observations and two variables X_{i1} and X_{i2}:

i	$Y_i(0)$	$Y_i(1)$	D_i	X_{i1}	X_{i2}	$\pi(\mathbf{X}_i)$
1	-	2	1	1	7	0.33
2	5	-	0	0	7	0.14
3	-	3	1	10	3	0.73
4	-	10	1	3	1	0.35
5	-	2	1	5	2	0.78
6	0	-	0	7	0	0.70

Ideally, we would match observations based on the exact propensity score, but as we can already see in this example, no two observations have the same $\pi(\mathbf{X})$. Instead, we have to approximate matches. This will create bias unless we assume that the distance will shrink as our number of observations grows.

It becomes a question of exactly how to construct the matches, and how many matches to choose. Do you pick just the closest neighbor? All neighbors within a fixed distance? This is a complicated problem that can affect inference and discussed in Abadie and Imbens [2008]. As an example, consider what happens if we impute based on the closest neighbor for each observation:

i	$Y_i(0)$	$Y_i(1)$	D_i	X_{i1}	X_{i2}	$\pi(\mathbf{X}_i)$
1	5	2	1	1	7	0.33
2	5	2	0	0	7	0.14
3	0	3	1	10	3	0.73
4	5	10	1	3	1	0.35
5	0	2	1	5	2	0.78
6	0	3	0	7	0	0.70

In this case, several units are used more than once as matches. Moreover, there are very close "ties": we picked unit $i = 1$ for unit 2, but unit 4 was very close as well. Why not pick that one, especially with $\pi(\mathbf{X})$ is noisy? This is a difficult problem to solve, and it is not clear that there is a single best answer.

The Horvitz-Thompson estimator works well but in smalll samples can be high variance if $\pi(\mathbf{X})$ is close to zero or one. This can be improved through the use of *stabilized weights*:

$$\hat{\tau}_{SIPW} = \frac{\frac{1}{n}\sum_i \frac{Y_i D_i}{\hat{\pi}(\mathbf{X}_i)}}{\frac{1}{n}\sum_i \frac{D_i}{\hat{\pi}(\mathbf{X}_i)}} - \frac{\frac{1}{n}\sum_i \frac{Y_i(1-D_i)}{1-\hat{\pi}(\mathbf{X})}}{\frac{1}{n}\sum_i \frac{(1-D_i)}{1-\hat{\pi}(\mathbf{X})}}$$

This estimator benefits by adjusting for unusually high or low values of $\pi(\mathbf{X})$ by constructing weights $w_i = \frac{\frac{D_i}{\hat{\pi}(\mathbf{X}_i)}}{\frac{1}{n}\sum_i \frac{D_i}{\hat{\pi}(\mathbf{X}_i)}}$ for the treated group.[5] Similar to the IPW, this is also an unbiased estimator of the ATE.

Comment 1 (True versus estimated propensity scores) *When an intervention is truly randomly assigned, the propensity score itself is known. However, in most non-experimental settings, the p-score is unknown and must be estimated. As we discussed above, we need to estimate $\pi(X)$ and it can be done in parametric, semi-parametric, or non-parametric fashion, depending on your assumptions about $\mathbf{X}$.*

It is useful to note that the model used to estimate the propensity score matters. For example, the linear probability model, which is commonly used for many binary outcomes, may predict probabilities for the propensity score that are outside the range of $[0, 1]$, thereby generating improper IPW estimates. The LPM will work if the model is fully saturated and non-parametric (e.g. a set of fully interacted dummies), but this is not always the case.

Another important result in this literature is that even if you know the true function $\pi(\mathbf{X})$, *you are better off using the estimated function than the true $\pi(\mathbf{X})$ [Hirano et al., 2003]. The intuition for this result is that the deviations from the "true" propensity score ($\hat{\pi}(\mathbf{X}) - \pi(\mathbf{X})$) are informative for the estimation of the treatment effects (a la extra moment restrictions in GMM)*

Contrasting linear regression with propensity scores

Say we have strong ignorability conditional on $\mathbf{X}$ and we run the following regression:

$$Y_i = \gamma_0 + D_i\beta + X_{i1}\gamma_1 + X_{i2}\gamma_2 + u_i. \tag{1}$$

How should we contrast this to the propensity score methods we used above?

We can revisit the Aronow and Miller [2019] example from Example 1, but instead of imputing for the missing counterfactual using matching, we impute using regression.

This tells us, roughly speaking, what this regression approach will assume for the missing counterfactuals. Notably, this approach will do well when the conditional expectation function for the outcome (e.g. $E(Y_i(0))$) is approximately linear in $\mathbf{X}$ and D_i – e.g. you are roughly correctly specified. Importantly, the approximation should not extrapolate too much across the support of $\mathbf{X}_i$.[6]

This is just another way to infer the missing data. However, a

[5] In the limit, recall that $\frac{1}{n}\sum_i \frac{D_i}{\hat{\pi}(\mathbf{X}_i)}$ should converge to 1 (since $E(\pi^{-1}(\mathbf{X}_i)D_i) = 1$), and hence the SIPW converges to the IPW.

[6] To concretely give an example of how this could be an issue: note that unit 3 is treated and is far out on the support of $\mathbf{X}_i$ (10,3). Imputing values for that unit's control requires extrapolating quite far out on the support of $\mathbf{X}_i$, which may be problematic unless the conditional mean is exactly correctly specified.

i	$Y_i(0)$	$Y_i(1)$	D_i	X_{i1}	X_{i2}	$\pi(\mathbf{X}_i)$
1	$\hat{\gamma}_0 + \hat{\gamma}_1 \cdot X_{i1} + \hat{\gamma}_2 \cdot X_{i2}$	2	1	1	7	0.33
2	5	$\hat{\gamma}_0 + \hat{\beta} + \hat{\gamma}_1 \cdot X_{i1} + \hat{\gamma}_2 \cdot X_{i2}$	0	0	7	0.14
3	$\hat{\gamma}_0 + \hat{\gamma}_1 \cdot X_{i1} + \hat{\gamma}_2 \cdot X_{i2}$	3	1	10	3	0.73
4	$\hat{\gamma}_0 + \hat{\gamma}_1 \cdot X_{i1} + \hat{\gamma}_2 \cdot X_{i2}$	10	1	3	1	0.35
5	$\hat{\gamma}_0 + \hat{\gamma}_1 \cdot X_{i1} + \hat{\gamma}_2 \cdot X_{i2}$	2	1	5	2	0.78
6	0	$\hat{\gamma}_0 + \hat{\beta} + \hat{\gamma}_1 \cdot X_{i1} + \hat{\gamma}_2 \cdot X_{i2}$	0	7	0	0.70

Table 1: Regression imputation in Aronow and Miller [2019] example

key issue is that if we just use OLS to estimate Equation (1), we will not necessarily recover the ATE with τ. Recall that when $\mathbf{X}$ is just a constant, then *tau* is the ATE. Once we condition on covaraites, however, the estimand recovered by τ changes. We will revisit this in our linear regression lectures, but the key fact is that OLS will recover an estimand which is a different weighted average of the CATEs – namely one that variance-weights the CATEs.

First, it's useful to recall what we are assuming with strong ignorability. Recall that we are assuming that there is a function $E(D_i|\mathbf{X}) \equiv \pi(\mathbf{X})$ that we can condition on such that there is no remaining correlation between D_i and the potential outcomes. However, we do not know the function $\pi(\mathbf{X})$. If we are additionally willing to assume that $\pi(\mathbf{X}) = X_{i1}\gamma_1 + X_{i2}\gamma_2$, then by Frisch-Waugh-Lovell, we can show that τ will recover a weighted combination of the CATEs.[7] Specifically, let $\tilde{D}_i = D_i - \pi(\mathbf{X}_i)$ and $\tau_i = Y_i(1) - Y_i(0)$. Then,

$$\beta_{OLS} = \frac{E(\tilde{D}_i D_i \tau_i)}{E(\tilde{D}_i^2)}. \tag{2}$$

As a result, the OLS coefficient will not necessarily estimate the ATE estimand. Instead, it will estimate a weighted average of the CATEs, where the weights are the variance of the treatment assignment conditional on $\mathbf{X}$. We will revisit this during our linear regression lectures.

As a result, there are two key caveats with this approach: first, we need to be careful about how we specify the regression function with our $\mathbf{X}$ variables. Especially if there is significant extrapolation across the support of $\mathbf{X}$.[8] Second, we need to be careful about how we interpret the coefficient on D_i in the regression. It will not necessarily be the ATE, but instead a variance-weighted average of the CATEs.

Consequentially, an important question for us as practitioners is which approach to use: IPW or regression? There are good reasons to like the IPW estimator for the ATE: Hirano et al. [2003] show that the IPW estimator is semiparametrically efficient when the propensity score is unknown.[9] But, linear regression is nice! It is straightforward to run, and easy to interpret. Moreover, it has been endorsed. From

[7] This discussion here follows Angrist [1998], and is generally inspired by Borusyak and Hull [2024].

[8] Note that in cases where we have two sets of fixed effects (e.g. age and location), we often just include the marginal fixed effects (e.g. age and location separately) and not the fully saturated specification(e.g. age × location). This is because the interaction fixed effects require too much of the data. However, there will consequentially be extrapolation that may be wrong.

[9] However, the IPW estimator can be high variance when the propensity score is close to zero or one (this is known as weak overlap). This can create issues generally with estimating the ATE, and linear regression avoids this issue. See Goldsmith-Pinkham et al. [2022].

Angrist and Pischke (2009):

> We believe regression should be the starting point for most empirical projects. This is not a theorem; undoubtedly, there are circumstances where propensity score matching provides more reliable estimates of average causal effects. The first reason we don't find ourselves on the propensity-score bandwagon is practical: there are many details to be filled in when implementing propensity-score matching - such as how to model the score and how to do inference - these details are not yet standardized. Different researchers might therefore reach very different conclusions, even when using the same data and covariates. Moreover, as we've seen with the Horvitz-Thompson estimands, there isn't very much theoretical daylight between regression and propensity-score weighting.

Comment 2 (Ex-Post Weights vs. Ex-Ante Weights)
Note that in Equation (2),

$$\phi_i(\mathbf{X}_i) = \tilde{D}_i D_i$$

is a function of $\mathbf{X}_i$ *and can be negative. Then that means for some CATE* ($E(\tau_i|\mathbf{X})$)*, there may be negative weights. In some special cases, this will not be the case (e.g.* $\pi(\mathbf{X})$ *is correctly specified and/or* $\mathbf{X}$ *is discrete and fully satuated). These negative weights can be problematic because the weighted TE could then reflect an effect that does not exist in the underlying population at all.*

However, as shown succinctly in Borusyak and Hull [2024], the expected *ex ante weights in a design-based approach are guaranteed to be positive:*

$$E(\phi_i|\mathbf{X}_i, \beta_i) = Var(D_i|\mathbf{X}_i, \beta_i) > 0.$$

This implies, intuitively, that all units with the same $\mathbf{X}_i$ *will have the same weight prior to treatment. This is a key difference between the design-based approach and the model-based approach. This same statement cannot be done in a context where we model* $E(Y_i(0))$ *because we are not allowing the treatment to be randomly allocated across units (by)*

When we start worrying about the propensity score, life gets more complicated. It forces us to think about overlap of covariates and balance, namely how comparable the treated and untreated groups are. This became a key issue around the seminal NSW paper.

NSW and Propensity score matching

There was a randomized intervention called the National Supported Work Demonstration (NSW), which was a temporary employment program to give work experience. A seminal paper by Lalonde [LaLonde, 1986] showed that a non-experimental analysis of this program would have given biased estimates compared to experimental approach.[10]

Dehejia and Wahba [2002, 1999] reanalyze this data using propensity score methods, and argue that these results are more similar to the experimental results, relative to the non-experimental results proposed by LaLonde [1986]. Moreover, using propensity scores provides a form of diagnostics on how comparable the treated and control groups are.

It is worth digging in to the Dehejia and Wahba [2002] paper to understand how they use propensity scores. Crucial to their approach is they included the lagged outcomes as covariates in their approach. As a consequence, they subsample the data to have two years of pre-treatment data. This is a key difference from the LaLonde [1986] approach, which used the full sample. Smith and Todd [2005] assess this approach, and argue two points: first, the specification itself for the propensity score is quite sensitive to the choice of included variables.[11] Second, they argue that the *subsampling* in Dehejia and Wahba [2002] predisposes to a group where the analysis is "easy." Moreover, in this case difference-in-differences works best because it removes the time-invariant heterogeneity across units.

Dehijia's response in Dehejia [2005] is "of course!" – the point of propensity score matching is to transparently highlight the assumptions in the data. More verbosely, he says:

> A judgment-free method for dealing with problems of sample selection bias is the Holy Grail of the evaluation literature, but this search reflects more the aspirations of researchers than any plausible reality. In practice, the best one can hope for is a method that works in an identifiable set of circumstances, and that is self-diagnostic in the sense that it raises a red flag if it is not functioning well. Propensity score methods are applicable when selection is based on variables that are observed. In the context of training programs, Dehejia and Wahba (1999, 2002), following on a suggestion from the training program literature (Ashenfelter, 1978; Ashenfelter and Card, 1985), suggest that two or more years of pre-treatment earnings are necessary. In terms of the self-diagnosis, the method and its associated sensitivity checks successfully identify the contexts in which it succeeds and those in which it does not succeed, at least for the NSW data.
>
> **Propensity score matching does not provide a silver-bullet, black-box technique that can estimate the treatment effect under all circumstances**; neither the developers of the technique nor Dehejia and Wahba

[10] "This comparison shows that many of the econometric proce- dures do not replicate the experimentally determined results."

[11] In other words, deciding which variables to include in **X** is important and can affect bias significantly. This echoes Leamer's critique on model specification in Leamer [1983].

> have claimed otherwise. However, with input and judgment from the researcher, it can be a useful and powerful tool. [Emphasis added]

What causes residual variation in treatment?

We initially motivated strong ignorablity under settings with random assignment or something approximating it. However, in many settings, the data researchers will use is exclusively observational and does not have random assignment. Despite that, they want to estimate a causal effect.

To quote Heckman et al. [1998]:

> Ironically, missing data give rise to the problem of causal inference, but missing data, i.e. the unobservables producing variation in D conditional on X, are also required to solve the problem of causal inference.

In other words, when we control for $\mathbf{X}_i$, there must be additional source of variation in D_i that we do not capture that drives differences in the choice of treatment, but is also unrelated to the potential outcomes.

Example 2 (Why do we need residual variation?)
Consider D to be a medical treatment selected by a doctor, with Y their subsequent health outcome. What would happen if D was perfectly predictable by $\mathbf{X}$: e.g., age of patient, the doctor's background, etc. In other words, if we know $\mathbf{X}$, then we know D.
In this setting where the $\mathbf{X}$ perfectly predict the treatment, is the effect of D on Y identifie after conditioning on $\mathbf{X}$? No. See this in two ways:

- *$Pr(D_i|\mathbf{X}_i) = 1$ or 0 and we fail overlap (and thus strong ignorability)*
- *$Y_i = D_i\tau + \mathbf{X}_i\gamma + \epsilon$ is our estimating equation, but $\mathbf{X}$ and D are perfectly collinear.*

To estimate the effect of D and Y, we need additional "exogeneous" variation.

A structural econometrician would describe the variation in D as driven by two pieces, V and X. Ideally, V is exogeneous. But what actually is V? Much of the time we don't know. This comes back to our research design question. Is there something "near-random" that caused a difference in treatment? Or if we choose to be pessimistic, if units are observably identical, but choose different outcomes, a purely rational model would suggest there are intrinsically different characteristics driving this decision. How will this bias our estimates (if at all)?

Consider Figure 1. There are many parts of $\pi(\mathbf{X})$ where there is lots of overlap between the treatment and control group. However, in some parts of the distribution there is significantly less, especially where $\pi(\mathbf{X}) < 0.5$. What does it mean to have so few treated units for the pscore less than 0.5? This suggests that ther are a small share of units who are both treated and look observably similar to a large set of the control group. It seems plasuible that these units might not be comparable. If we choose not to use the units at the "extremes", e.g. by selecting on only propensity scores between 0.5 and 0.75, what would that imply about our model estimates? This would be targeting a very particular estimand that may or may not be of interest.

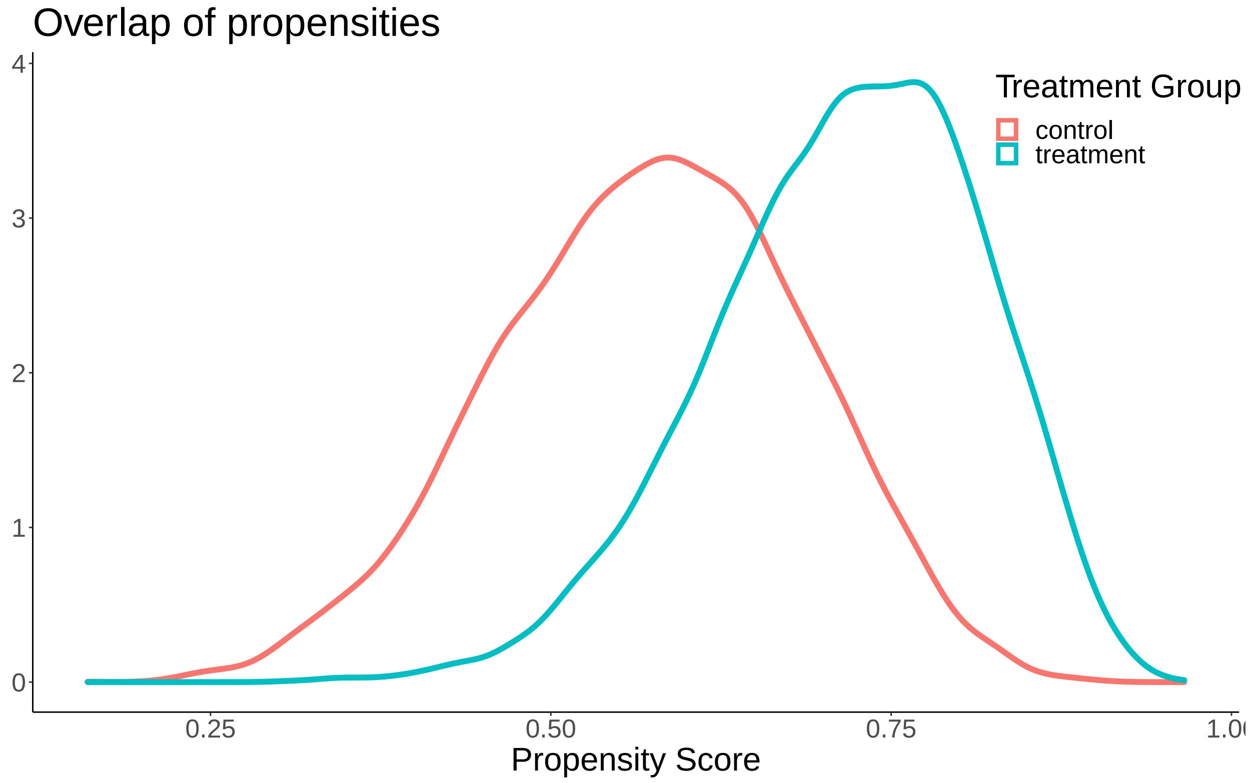


Figure 1: Overlap in the propensity scores for example treated and control populations

As we go forward in thinking about random variation in D_i, it is convenient to consider the following economic model (from Heckman [1997]):

$$Y_i(0) = g(X_i, D_i = 0) + U_{i0}$$
$$Y_i(1) = g(X_i, D_i = 1) + U_{i1}$$
$$Y_i = g(X_i, 0) + D_i \left(\underbrace{g(X_i, 1) - g(X_i, 0)}_{\text{Average Population Gain}} + \underbrace{U_{i1} - U_{i0}}_{\text{idiosyncratic gain}} \right) + U_{i0}$$

Now, we consider what drives the decision making for D_i:

$$D_i = 1((Y_i(1) - Y_i(0)) + \kappa + V_i > 0)$$

In other words, when the value is sufficiently high (above some overall + idiosyncratic cost $\kappa + V_i$), I choose to take the program. This creates obvious correlation between D_i and $(Y_i(0), Y_i(1))$.

WIth this setup, we can consider under what settings controlling for $\mathbf{X}$ will be sufficient to recover a causal estimand. The easiest is when there are constant effects, e.g. $U_{i1} - U_{i0} = 0$ for everyone. In this case, random variation in V_i is what drives takeup, and is unrelated to the outcome. This makes life easy for us, but is not very interesting and also means there is no underlying heterogeneity in the treatment effect beyond what we observe in the characteristics.

The other case is when the *expected* gains to the program is the same for everyone ($E(U_{i1} - U_{i0}|X_i) = 0$), perhaps because of lack of information on the part of the individuals'. Then, while there may be ex post differences in treatment effects, they are not ex ante anticipated by the individuals and consequentially selected on.

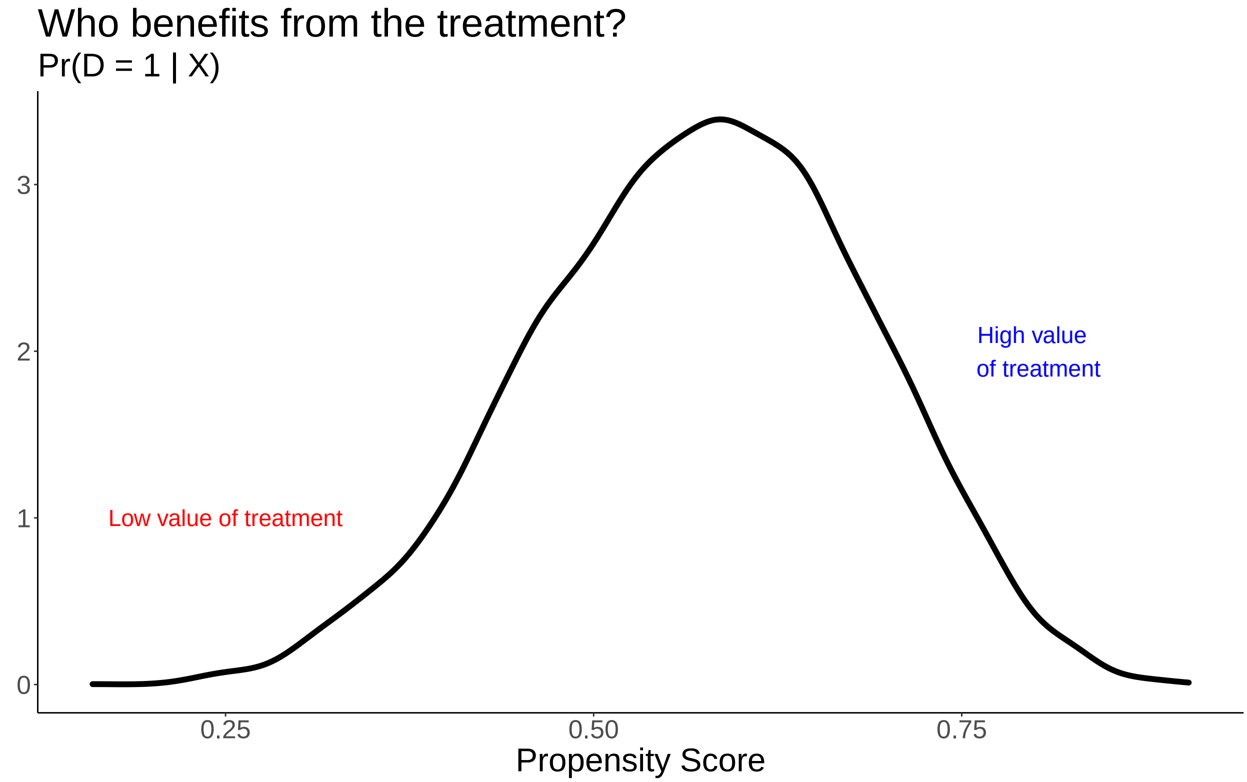
The propensity score in this model is:

$$Pr(D_i = 1|X_i) = Pr\left(g(X_i, 1) - g(X_i, 0) + \kappa + (U_{i1} - U_{i0}) > V_i\right)$$

This gives us a framework to consider the economic returns to individuals to take a program, as in **??** . What would it take to switch them into the program?

1. Lack of choice: not always available
2. Large incentive: expensive
3. High personal returns: selects into a particular type of person.

It is useful to remember this graph when considering how to induce participation. Some folks may just not want to participate. This could be perceptions on the returns (e.g. $Y(1) - Y(0)$), rightly or wrongly, but as a consequence, they will be expensive to move. The estimand of interest will be considering parts of this distribution, and hence it si important to consider this when thinking about external validity.



Going forward, it is helpful to consider the propensity score as an index of valuation. We are hence looking for things that vary individuals' valuation and do not correlate with the potential outcome. In design-based work in economics, this is often referred to as an *instrument*, and is a crucial part of the design-based research.

References

Alberto Abadie and Guido W Imbens. On the failure of the bootstrap for matching estimators. *Econometrica*, 76(6):1537–1557, 2008.

Joshua D. Angrist. Estimating the labor market impact of voluntary military service using social security data on military applicants. *Econometrica*, 66(2):249–288, 1998. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/2998558.

Peter M Aronow and Benjamin T Miller. *Foundations of agnostic statistics*. Cambridge University Press, 2019.

Kirill Borusyak and Peter Hull. Negative weights are no concern in design-based specifications. Technical report, National Bureau of Economic Research, Inc, 2024.

Rajeev Dehejia. Practical propensity score matching: a reply to smith and todd. *Journal of econometrics*, 125(1-2):355–364, 2005.

Rajeev H Dehejia and Sadek Wahba. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American statistical Association*, 94(448):1053–1062, 1999.

Rajeev H Dehejia and Sadek Wahba. Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and statistics*, 84(1):151–161, 2002.

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

James Heckman. Instrumental variables: A study of implicit behavioral assumptions used in making program evaluations. *Journal of human resources*, pages 441–462, 1997.

James J Heckman, Hidehiko Ichimura, and Petra Todd. Matching as an econometric evaluation estimator. *The review of economic studies*, 65(2):261–294, 1998.

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

Robert J LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, pages 604–620, 1986.

Edward E Leamer. Let's take the con out of econometrics. *The American Economic Review*, 73(1):31–43, 1983.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Jeffrey A Smith and Petra E Todd. Does matching overcome lalonde's critique of nonexperimental estimators? *Journal of econometrics*, 125 (1-2):305–353, 2005.

Lecture 4 - Interference, Spillovers and Dynamics

Paul Goldsmith-Pinkham

January 25, 2024

This lecture note will discuss what it means to relax the following assumptions from the previous lecture:

1. Binary scalar treatment
2. Single time period (e.g. one treatment within the person)
3. SUTVA – Stable Unit Treatment Value Assignment

Multivalued treatments

So far, our discussion of treatment effects has focused on single binary treatments. This made life very easy, but we have a lot of other more complex settings. We'll consider a few different cases. First, a multi-valued treatment. Then, we'll consider a continuous treatment. Finally, we'll consider an unordered multi-valued treatment.

Discrete multi-valued treatment

Let's start with a discrete, multi-valued treatment to start: $D_i \in \mathcal{D} = \{0, 1, \ldots, d\}$. This captures a simple setting like "what is the impact of 0, 1, 2, or 3 children on labor force participation?" In this setting, we can easily shift the scale up and down ("what is the impact of 5, 10, 15, or 20 minutes on a task") but the order and spacing may matter, depending on how we choose to parameterize and estimate the treatment effect.

In this setting, how should we consider the treatment effects? First, what should we consider the "control"? We could consider the control to be the lowest value of the treatment, but that depends a bit on the context. For example, if we are considering the impact of 0, 1, 2, or 3 children on labor force participation, we might consider the control to be 0 children. However, if we are considering the impact of 5, 10, 15, or 20 minutes on a task, we might consider the control to be whatever the status quo was.

Formally, we define the potential outcome for any $d \in \mathcal{D}$ as $Y_i(d)$, and we consider the individual and average treatment effect difference between d and d' as:

$$\begin{aligned} \tau_i(d, d') &= Y_i(d) - Y_i(d') \\ E(\tau_i(d, d')) &= E(Y_i(d) - Y_i(d')). \end{aligned}$$

If strong ignorability holds, then this is also identified by simply conditioning on each observed value:[1]

$$E(\tau_i(d, d')) = E(Y_i|D_i = d) - E(Y_i|D_i = d').$$

This type of estimation is non-parametric in nature: we've assumed no functional form between the treatment and the potential outcome. A consequence of that, just like in the case with many covariates $\mathbf{X}$, is that it requires a lot more data to provide precise estimates. If we wanted to consider the CATE:

$$E(\tau_i(d, d')|\mathbf{X} = x) = E(Y_i|D_i = d, \mathbf{X} = x) - E(Y_i|D_i = d', \mathbf{X} = x),$$

then we'll need to condition on treatment catgories within each cell, which can be very data hungry, and less precise.

Often, instead of estimating the effect for every point separately, we will postulate a model for the potential outcomes:

$$Y_i(d) = Y_i(0) + \tau_i d.$$

Notice that in this case, this implies that for all d, d' pairs

$$\tau_i(d, d') = \tau_i,$$

which is the slope parameter. Hence, estimation can be made more precise by pooling all of these estimands together into a single estimand $E(\tau_i)$.[2]

Example 1
Consider the following simulated data, where the true effect is linear and simulated such that $E(\tau_i(d, d') = d' - d$ and strong ignorability holds.

Each dot in Figure 1 is the estimated mean at the point, and we find a positive treatment effect. Imposing the model helps a lot *compared to non-parametric form. To see this, consider the treatment effect comparing d to $d - 1$ in Figure 2*

The direction of the effect is much more ambiguous. This is a common tradeoff in estimation: imposing a model can help with precision, but can also lead to bias if the model is misspecified.

How should we consider these functional forms once we include controls? To see what I mean, consider the same context, but we now assume strong ignorability conditional on $\mathbf{X}$. Then, we would need to estimate the slope τ_i for each value of $\mathbf{X}$. How would that map over to a linear regression model? The simplest version would be one where the heterogeneity, τ_i, is uncorrelated with $\mathbf{X}$. Then, one

[1] The overlap condition in strong ignorability with multiple treatments is more complicated, but effectively entails that for any $\mathbf{X}$, there are observations for every d in $\mathcal{D}$.

[2] Other, more flexible parametric forms for $Y_i(d)$ could be chosen as well. The insights will carry through so long as the functional form is finite-dimensional.

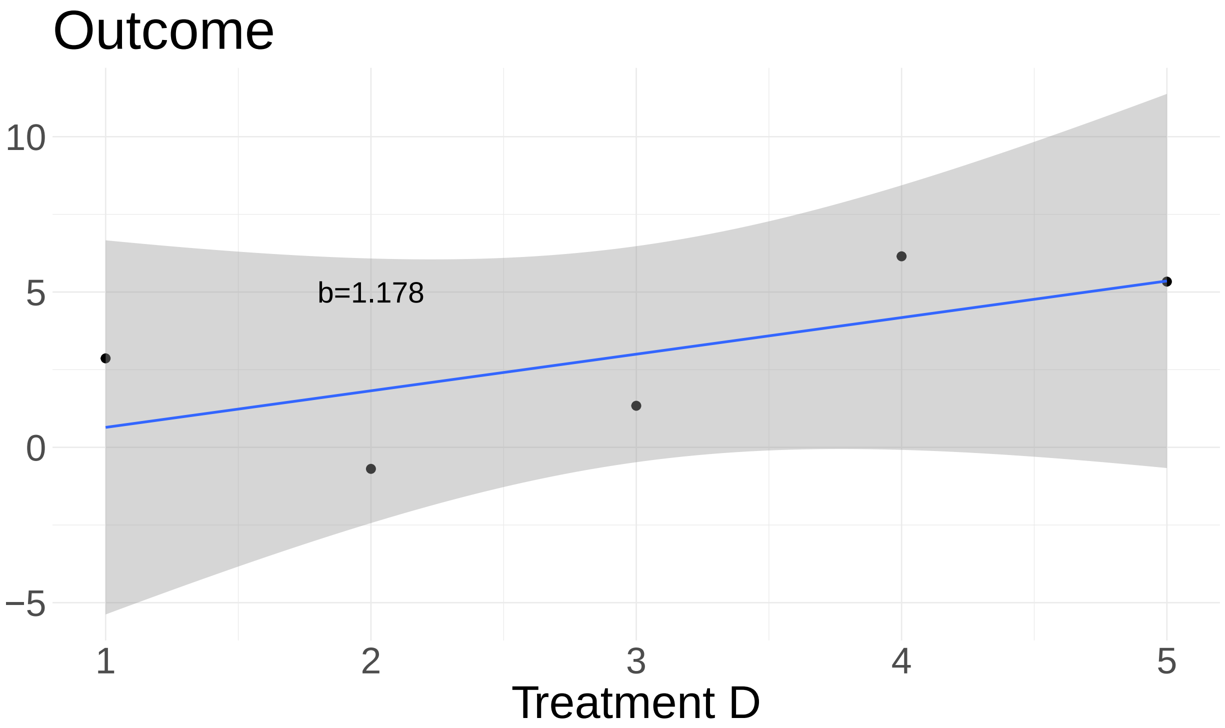


Figure 1: Linear effects estimated in simulated data with a true linear model for Example 1

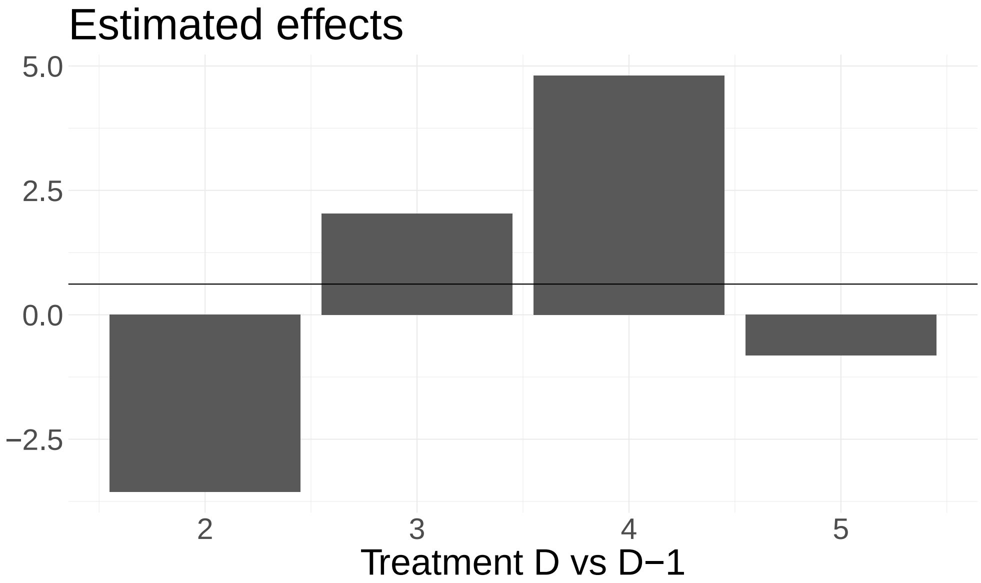


Figure 2: Non-parametric effects estimated in simulated data with a true linear model for Example 1

way to estimate the ATE is to assume that $E(D_i|\mathbf{X}) = \mathbf{X}_i\beta$ (e.g. the propensity score is linear in $\mathbf{X}$), we could estimate $E(\tau_i)$ by simply running the following regression:

$$Y_i = \alpha + D_i\tau + \mathbf{X}_i\beta + \epsilon_i \tag{1}$$

and using τ as our estimate of $E(\tau_i)$. But, if τ_i is not uncorrelated with $\mathbf{X}$, then to estimate the ATE we would need to estimate the slope for each value of $\mathbf{X}$, and pool separately. See Comment 1 for more details.

Comment 1

It is worth thinking about why Equation (1) will correctly estimate the ATE in this setting. To do this, let $E^*(D_i|\mathbf{X}_i)$ denote the best linear predictor of D_i conditional on $\mathbf{X}_i$. Now, note that by Frisch-Waugh-Lovell,

$$\tilde{Y}_i = \tilde{\alpha} + \tilde{D}_i\tau + u_i, \tag{2}$$

where $\tilde{Y}_i = Y_i - E^*(Y_i|\mathbf{X}_i)$ and $\tilde{D}_i = D_i - E^*(D_i|\mathbf{X}_i)$. Then,

$$\begin{aligned}\tau &= \frac{E(\tilde{D}_iY_i)}{E(\tilde{D}_i^2)}\\ &= \frac{E(\tilde{D}_iY_i(0))}{E(\tilde{D}_i^2)} + \frac{E(\tilde{D}_iD_i\tau)}{E(\tilde{D}_i^2)}\end{aligned}$$

The first term is zero because the residual $\tilde{D}_i$ is mean independent of $\mathbf{X}$ by the linearity of $E(D_i|\mathbf{X})$. Therefore,

$$E(\tilde{D}_iY_i(0)) = E(E(\tilde{D}_iY_i(0)|\mathbf{X})) = E(E(\tilde{D}_i|\mathbf{X})E(Y_i(0)|\mathbf{X})) = 0.$$

Now note the second term by similar arugments:

$$\tau = \frac{E(\tilde{D}_iD_i\tau)}{E(\tilde{D}_i^2)} = \frac{E(Var(D_i|\mathbf{X})E(\tau|\mathbf{X}))}{E(Var(D_i|\mathbf{X}))}.$$

But, since we assumed $E(\tau|\mathbf{X}) = E(\tau)$, this is just the ATE. If there is correlation of τ with $\mathbf{X}$, we will get a different estimand.

Discussion Questions 1

Under what assumptions about $E(Y_i(0)|\mathbf{X}_i)$, instead of $E(D_i|\mathbf{X}_i)$, could we estimate the ATE using Equation (1)?

1. *Hint: think about a linear model for $E(Y_i(0))$ that is linear in $\mathbf{X}_i$.*
2. *Second hint: the estimand could also be written as*

$$\tau = \frac{E(\tilde{D}_i \tilde{Y}_i)}{E(\tilde{D}_i^2)}$$

Continuous valued treatment

In many cases, the jump from discrete ordered treatments to continuous valued treatments is not large. Often, it just has to do with how many repeated observations we have of the same treatment; if each treatment value is unique, we're more likely to treat it as continuous. None of what we discussed above changes, except that direct non-parametric estimation becomes infeasible.

Instead, to do non-parametric estimation we'll need to make other assumptions and use other methods, like kernel regression or local linear regression. We'll discuss these in future classes, but the key point is that we will want to make some amount of smoothness assumptions the effect of the treatment on the potential outcome.

Instead of non-parametric estimation, it is also reaosnable to assume a functional form, as above, and proceed from there. Then everything is exactly the same.

Unordered multi-valued treatment

Finally, we might have a setting where the treatment is unordered. For example, we might consider the impact of different CEOs on firms' performance. In this case, we can't assume any particular ordering of the treatment, and it's not clear how to presume a functional form. Instead, there is a set of K treatments in $\mathcal{D}$, and we can consider a set of different contrasts between them: $E(\tau_i(d, d'))$ for all $d, d' \in \mathcal{D}$.

A straightforward special case would be to consider a *factorial* design: a randomized treatment where two treatments are cross-randomized, such that an individual can receive either no treatment, treatment 1, treatment 2, or both. Then, our potential outcomes look like Table 1.

Given this, we have a number of potential estimands to consider. For example, we could consider the average treatment effect of treatment 1, but we would need to make a decision on what to do about the individuals who received both treatments. If the treatments interact in some way, then the average treatment effect of treatment 1 is not well-defined. Instead, we might consider the average treatment effect of treatment 1 for those who received treatment 2 ($E(Y_i(1,1) - Y_i(0,1))$), and the average treatment effect of treatment 1 for those who did not receive treatment 2 ($E(Y_i(1,0) - Y_i(0,0))$).

$Y_i(\mathbf{D_i})$	$D_{1i} = 0$	$D_{1i} = 1$
$D_{2i} = 0$	$Y_i(0,0)$	$Y_i(1,0)$
$D_{2i} = 1$	$Y_i(0,1)$	$Y_i(1,1)$

Table 1: Potential outcomes for a factorial design

Of course, when data is sparse, we might want to do more with this, and just pool all of the data together: $E(Y_i(1, D_{i2}) - Y_i(0, D_{i2}))$. This is a reasonable estimand if we believe that the treatment effects are constant across the different levels of the other treatment, but if there are interactions, the external validity of this estimate will be suspect.[3] See Banerjee et al. [2021] for a very interesting discussion on how to think about these types of estimands in the context of factorial designs when there are many treatments.

Factorial designs are the simplest case to consider, because the treatments are cross-randomized for many binary treatments. Often, we have just an ordered set of treatments. In this case, the same logic applies, but we have to be more careful about how we define the estimands. For example, if we consider the impact of different CEOs on firm performance, how do we define a “control”? It is often not obvious, and we need to be careful about how we define the estimands. We may instead focus on the conditional means for each treatment, and then consider the full distribution of effects. This is the type of consideration in work thinking about place-based effects, for example, such as Chetty and Hendren [2018].[4]

Estimating these effects seem like they should be straightforward extensions of the binary treatment case. However, the partial linear regression model fails us in this case. The variation in the propensity score across strata (controls) combined with heterogeneity in the treatment effects across strata will lead to contamination bias in the linear regression model. See Goldsmith-Pinkham et al. [2022] for more details, which we will revisit in the linear regression lectures.

Another issue that can arise is when the treatments are not cross-randomized, but instead are correlated on one another. A simple example of this is sequenced treatments: e.g. treatment 2 is only given after treatment 1, and only a subset of individuals receive treatment 1. See Figure 3 for an example of this. In this case, it is not possible to identify the effect of D_2 separately from D_1: $E(Y_i(0,1) - Y_i(0,0))$ is not identified because $E(Y_i(0,1))$ is never observed. This rarely

[3] Think about why this is the case, if it’s not clear.

[4] It is interesting to think about these unordered treatments can sometimes be projected into continuous scalar measures. For example, when considering the impact of a CEO on a firm, you might measure a CEO’s experience, and project the overall CEO’s effect onto experience to capture a continuous measure of the CEO’s effect. If you are additionally willing to assume that the effect of experience is the *only* channel controlling the CEO’s effect, then a more efficient procedure would use experience as the treatment effect, instead of the CEO. But that may not be a reasonable assumption. This issue arises when considering the effect of judges as in Arnold et al. [2022].

happens in many cross-sectional settings, but is quite common in dynamic settings (our next topic).

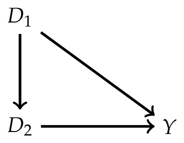


Figure 3: Correlated treatments

Treatment dynamics

We will briefly discuss the impact of treatments over time to set the stage for our study of panel data later in the class. Consider a setting where we now observe T time periods for a unit: $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{iT})$. Now, for each time period, there is a treatment D_{it}. It would be convenient to simply consider $Y_{it}(D_{it})$ as the potential outcome for an individual i in period t, but that would be very restrictive: it would assume that only the period t treatment affects the outcome in period t. A more general form would define a vector $\mathbf{D}_i = (D_{i1}, D_{i2}, \ldots, D_{iT})$, and define the potential outcome in period t as $Y_{it}(\mathbf{D}_i)$. In this case, however, we are perhaps too general: this allows for treatments in the future to affect current outcomes, which may be too strong.[5]

There are a large number of ways to simplify these potential outcomes. One simple way would be to restrict treatments to only affect outcomes in the future, and not the past. This is often referred to as the "no anticipation" assumption. The second is to assume that treatments will only turn on once: this allows the researcher to only consider the adoption date as the relevant period.[6]

As you can see, things become much more complex as soon as you allow for dynamic effects. In order to make progress, it will often be necessary to make restrictions on teh dynamics to make the estimands identified. We will discuss these in more detail when we discuss difference-in-differences.

The SUTVA hits the fan

In the discussion so far, the "interference" between treatments just comes from having multiple treatments to worry about, or from spillover across time. However, there are many other ways that treatments can interfere with one another. For example, what if treatments spill across units? What if the treatment of one unit affects the potential outcomes of another unit?

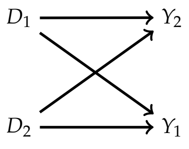


Figure 4: Interference between units

Recall the key assumption of Stable Unit Treatment Value Assumption (SUTVA): the potential outcomes of a unit do not vary with the treatment of other units. When could this be violated?

So many places

Why does failure of SUTVA create an issue? Recall our discussion regarding marginal estimands when there were multiple treatments:

[5] This is often referred to as "anticipatory effects" in the literature.

[6] This is common in the staggered difference-in-difference setting.

even with random assignment, the estimates effect will be contaminated by others' treatment status, thereby leading to estimates that are not informative for the policy maker.

This type of problem is generally referred to as "interference." It is challenging for identification, estimation and inference. For now, we'll focus on identification. I will flag three versions of this problem:

1. Social interactons and peer effects

2. Spatial spillovers

3. Economic interactions – budget constraints, etc.

All these problems are versions of violation of SUTVA. With a clean, well-identified experiment, it is still possible to identify interesting estimands, but we may have to substantially modify our traditional estimators or make strong assumptions to make progress. One way to view this fact is that our original setting — SUTVA, binary treatment and a single time period — is a very special (and somewhat unrealistic) case.

Social interactions and peer effects

> A variety of terms in common use connote endogenous social effects, wherein the propensity of an individual to behave in some way varies with the prevalence of that behaviour in some reference group containing the individual. These effects may, depending on the context, be called "social norms", "peer influences", "neighbourhood effects", "conformity", "imitation", "contagion", "epidemics", "bandwagons", "herd behaviour", "social interactions", or "interdependent preferences".
> — Manski [1993]

Manski (1993) spawned a huge literature, much of which focused on the linear-in-means model.[7]. An inherent issue, in my view, is that many empirical papers jumped to this construction immediately. They did not have a structural interpretation in mind, but were instead interested in testing for the *statistical* presence of spillovers across individuals.

[7] There are theoretical models microfounding a linear-in-means outcome model, which typically involve some kind of quadratic cost to deviating from the group. See Shue [2013] for an example

Comment 2 (Historical context on peer effects)
Manski [1993] focused on a linear-in-means structural equation

$$Y = \underbrace{\beta E(Y|g)}_{endogeneous} + \underbrace{\gamma_1 E(X|g)}_{exogeneous} + \gamma_2 X$$
$$+ \underbrace{\gamma_3 g}_{contextual} + u.$$

Peers were not well-defined in the model, but empirically, were usually groups like classrooms or clubs. What is important to note about this model is that it is a structural *model of the outcome, Y. The reduced form is*

$$Y = \gamma_1/(1-\beta)E(X|g) + (\gamma_2/(1-\beta))X$$
$$+ (\gamma_3/(1-\beta))g + \tilde{u},$$

which is estimable under special exogeneity assumptions on X, g and E(X|g).

An innovation in this literature was to start using network data to define the group structure. Bramoullé et al. [2009] was a key paper in this literature, that reframed the Manski linear-in-means model to

$$Y = \beta AY + \gamma_1 AX + \gamma_2 X + \epsilon_i,$$
$$Y = (I-\beta A)^{-1}\gamma_1 AX + (I-\beta A)^{-1}\gamma_2 X + (I-\beta A)^{-1}\epsilon_i$$

where A was an n × n matrix of individuals' connections. This was still a structural model, but allowed for richer data and more easily identified the effect of peers.

Given that most researchers studying peer effects were not initially motivated by a structural model, it seems more natural to initially take a statistical approach to the problem. Namely, we would like to identify the effect of spillovers across units. How can we approach this problem using the tools we've developed so far?

Given n individuals, for person i, how much interference can we allow? What types?

$$Y_i(D_1, D_2, \ldots, D_n)$$

is far more extreme than

$$Y_i(D_i, A\mathbf{D}_n).$$

This question is analogous to our setting with treatment dynamics: how much spillover should we allow? SUTVA is complicated by the fact that there is no natural "no anticipation" condition due to the natural flow of time. As you might expect, there is no "one solution" in this setting. Certain restrictions need to be made to identify some estimands.

Manski [2013] is a very nice discussion of this in a *very* high-level way. One key assumption he highlights is that "anonymity" of treatment spillover is a very important assumption. This implies that if I have peers who are treated, it does not matter *which* of those peers are treated – the impact on me is identical. This is a very strong assumption, but it is a necessary one to identify the effect of peers. If each peer's effect is allowed to be unique, then the effect of peers is not identified, since there is no way to separate out the treatment's spillover effect from the effect of the peer itself.[8]

A key question to keep in mind when considering spillovers: are you attempting to estimate the *spillover* effect, or are you attempting to identify individual ATE in the presence of spillovers? These are very different estimands, and require different assumptions to identify. For the purposes of external validity, the latter is really only relevant if the context you apply the treatment in would have limited spillovers as well.

We now briefly discuss two papers in this space to give intuition.

Aronow and Samii [2017] is a lynchpin paper in this setting that provides a framework for thinking about estimation and identification under general forms ofb interference. They use design-based inference, and consider the following generalized mapping.

Definition 1
For any generalized vector of interventions, $\mathbf{D}_n$, there's an experimental design which assigns probabilities over $\mathbf{D}_n$. There is then an exposure *mapping $f(\mathbf{D}_n, \theta_i)$ from these vectors to a treatment for an individual, which includes traits of an individual, θ_i (e.g. their network location) and the treatment vector, and maps it to an exposure outcome.*

This exposure mapping does two things. First, it makes restrictions on types of interactions (e.g. who can affect you and what type of effect it is).[9] Second, it maps the experimental design to a propensity score of the exposure treatment. This allows the use of Horvitz-Thompson estimators.

So where are the bodies buried in this method? You have to have a correctly defined exposure mapping, and you have to have a correctly defined experimental design. In the case of a randomized experiemnt, the latter is straightforward, but the former is not, and typically needs to be motivated by theory, or asssessed for robustness. This is an active literature.

[8] It might be doable in some networks, but it would be very challenging to do so, and require exogeneous network connections.

[9] Concretely: consider a network of peers affecting you. Is it the sum of your connected individuals in your network? Any exposure at all? Does it matter who in your network exposes you?

Athey et al. [2018] studies null hypothesis tests in networks under intereference. A key feature that this paper adds: testing specific types of analysis by creating "artificial" experiments. This paper is particularly powerful because it allows for testing in settings where there is uncertainty about the exposure mapping, and gives a framework for thinking about testing in a single network.

Comment 3
When thinking about experiments in networks (and other settings), the structure of spillovers is very important. It is extraordinarily helpful to identify settings where there are zero *spillovers. Having units (such as villages, roommate pairs, etc.) that are isolated from one another is a very helpful way to identify the effect of the treatments. If we permute the treaments across these groups, then we can assess the spillover effects in a very clean way.*

If, instead, we have only a single network, then we need to make strong assumptions about the structure of the network to identify the spillover effects. Namely, we need to have a well-defined exposure mapping that asserts that some units are sufficiently independent from others to serve as control units.

It is already very hard to do research on spillovers. Make sure to not ignore the difficult identification challenges and assumptions that you'll need to make. If you need a model, that's great! But often you are just interested in starting from a statistical perspective, which suggests you shoudl focus on a design-based approach as in Aronow and Samii [2017].

Identification of Endogenous Social Effects: The Reflection Problem
CHARLES F. MANSKI
University of Wisconsin-Madison
First version received December 1991; final version accepted December 1992 (Eds.)
Manski (1993)
Identification of treatment response with social interactions
CHARLES F. MANSKI
epartment of Economics and Institute for Policy Research, Northwestern University, 2
Sheridan Road, Evanston, IL 60208 USA
E-mail: cfmanski@northwestern.edu
First version received: February 2011; final version accepted: November 2011
Manski (2013)
imgflip.com

Figure 5: My views on social interactions summed up

Spatial Spillovers

Much of the spatial literature has sat in the same literature as social interactions. Distance on a network graph can be viewed as a similar distance metric to geographic (or economic) distance.Similar A matrix, and consequentially similar structural models are proposed.

The Aronow and Samii [2017] setting allows for this as well. From an identification standpoint, there isnothing deeply different here relative to networks, except that distance is potentially more continuous / complex. When we revisit simulated instruments, we will discuss some interesting implications raised by Borusyak and Hull [2020].

Economic interactions

Consider the following simple experiment – I give one half of people in the economy checks for $2000 dollars. I then study the impact of these checks on their consumption. Why might the effects be different than if I had run this experiment on a small share of individuals?

The economic spillovers coming through budget constraints are hugely important, but also deeply challenging as well. They require, often, modeling assumptions about spillovers. I will discuss two examples from the literature to give a flavor of the issues and solutions.

Chodorow-Reich [2019] studies the impact of fiscal stimulus on local employment. The key identification strategy is to use cross-region incidence of fiscal stimulus to identify multipliers on local employment. The paper argues that cross-region evidence bounds the estimand of interest, the impact of a *national* stimulus, from below.

> Drawing on theoretical explorations, I argue that the typical empirical cross-sectional multiplier study provides a rough lower bound for a particular, policy-relevant type of national multiplier, the closed economy, no-monetary-policy-response, deficit-financed multiplier. The lower bound reflects the high openness of local regions, while the "rough" accounts for the small effects of outside financing common in cross-sectional studies.

Sraer and Thesmar [2023] use cross-firm experiment to influence the allocation of credit. Some firms got lots more credit! Some did not. How to aggregate up this affect? E.g. the policy effect is estimated by differencing the impact of the change on those who were more directly exposed vs. not – however, this doesn't tell us about the aggregate impact on the economy. The paper argues, using economic theory, that these issues can be safely ignored under certain assumptions.

> Our paper bridges these two approaches. We offer a method to measure allocative efficiency in a (quasi-) experimental settings. This method works as follows. An econometrician observes firm-level data in an economy where a (quasi-) natural experiment has taken place. This experiment changes the set of frictions faced by treated firms while leaving control firms unaffected. Under the appropriate identifying assumption, the econometrician can estimate the causal effect of the experimenton firm-level outcomes, using classic difference-in-difference estimators. Standard policy evaluations typically estimate treatment effects on firm size or employment. However, these treatment effects alone cannot speak to allocative efficiency. To do so, we show that the econometrician needs to estimate treatment effects on the distribution of log marginal products of capital (lMRPKs). These estimates can then be injected in a simple aggregation formula to answer two simple questions: (i) how much did the actual policy change contribute to changes in aggregate efficiency (ex post evaluation)? (ii) how would aggregate efficiency have changed if the policy had been extended to all firms in the economy (scale-up)?

References

David Arnold, Will Dobbie, and Peter Hull. Measuring racial discrimination in bail decisions. *American Economic Review*, 112(9): 2992–3038, 2022.

Peter M Aronow and Cyrus Samii. Estimating average causal effects under general interference, with application to a social network experiment. *Annals of Applied Statistics*, 11(4):1912–1947, 2017.

Susan Athey, Dean Eckles, and Guido W Imbens. Exact p-values for network interference. *Journal of the American Statistical Association*, 113(521):230–240, 2018.

Abhijit Banerjee, Arun G Chandrasekhar, Suresh Dalpath, Esther Duflo, John Floretta, Matthew O Jackson, Harini Kannan, Francine N Loza, Anirudh Sankar, Anna Schrimpf, et al. Selecting the most effective nudge: Evidence from a large-scale experiment on immunization. Technical report, National Bureau of Economic Research, 2021.

Kirill Borusyak and Peter Hull. Non-random exposure to exogenous shocks: Theory and applications. Technical report, National Bureau of Economic Research, 2020.

Yann Bramoullé, Habiba Djebbari, and Bernard Fortin. Identification of peer effects through social networks. *Journal of econometrics*, 150 (1):41–55, 2009.

Raj Chetty and Nathaniel Hendren. The impacts of neighborhoods on intergenerational mobility ii: County-level estimates. *The Quarterly Journal of Economics*, 133(3):1163–1228, 2018.

Gabriel Chodorow-Reich. Geographic cross-sectional fiscal spending multipliers: What have we learned? *American Economic Journal: Economic Policy*, 11(2):1–34, 2019.

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

Charles F Manski. Identification of endogenous social effects: The reflection problem. *The review of economic studies*, 60(3):531–542, 1993.

Charles F Manski. Identification of treatment response with social interactions. *The Econometrics Journal*, 16(1):S1–S23, 2013.

Kelly Shue. Executive networks and firm policies: Evidence from the random assignment of mba peers. *The Review of Financial Studies*, 26 (6):1401–1442, 2013.

David Sraer and David Thesmar. How to use natural experiments to estimate misallocation. *American Economic Review*, 113(4):906–938, 2023.

Lecture 5 – Linear Regression 1 – Inference

Paul Goldsmith-Pinkham

January 31, 2024

This lecture note will discuss focus on the simple linear model, and studying various cases for understanding inference. So far we've focused on *identification* – e.g. what estimands can we know from the data generating process? Now, given estimators for these estimands, we want to discuss uncertainty and inference.

In this lecture note, we'll be dancing around a fundamental tension. Much of the existing literature and approach to inference is model-based, and as a result, you may feel a bit of a bait-and-switch: I promised a lot of design-based discussion, and all of a sudden we're back in the old world.[1]

The reason for this is that the model-based world is a useful starting point for understanding the basic ideas of inference. A lot of important statistical ideas will show up in this setting, and it's also the world that most of the literature is written in, so it's important to understand the basic ideas. Overall, just think of these different approaches as different tools in your toolbox.

Review: random sampling and linear regression

To fix notation, I want to do a refresher on notation and concepts about random sampling. This section is heavily based on review of materials from Gary Chamberlain, to which I am very grateful.[2] Conceptually, we have considered random variables (Y, X, D) from a joint distribution F:

$$(Y, X, D) \sim F. \qquad (1)$$

Now, we will formalize the concept of the data we observe that approximates the full population. We'll consider a sample of size n, where the ith draw of the data gives us the variables (Y_i, X_i, D_i).[3] We will initially consider random sampling where each draw is independent and identically distributed (i.i.d.):

$$(Y_i, X_i, D_i) \overset{i.i.d.}{\sim} F. \qquad (2)$$

Notationally, I'll often group $W_i = (X_i, D_i)$[4] in order to focus on a single set of non-outcome variables. The random sampling happens jointly – we make no distinctions between Y and W in the sampling process. We make that distinction later when we consider the conditional expectation of Y_i conditional on W_i. We will then stack the data $Y_n = (Y_1, \ldots, Y_n)$, $W_n = (W_1, \ldots, W_n)$.

[1] Recall that when I talk about design-based inference, I am thinking about treating the potential outcomes as fixed $(Y(1), Y(0))$ and the treatment assignment D as random. Model-based inference, in the context of lecture note, is considering the model $Y = X\beta + D\tau + \epsilon$, and then thinking about the random sampling of the (Y, X, D) creating uncertainty in the estimates. In other words, it's about the variation in $\epsilon | X, D$.

[2] Gary Chamberlain was an extraordinary econometrician and former teacher of mine who had an amazing set of lecture notes that I got permissiont to post online.

[3] In panel settings, with T observations, it is easy to consider Y_i being a vector of $Y_{i1}, Y_{i2}, \ldots, Y_{iT}$, and similarly for the other variables. Then, these vectors will be treated as a unit.

[4] When I need to make a distinction, X will usually denote controls and D will be the causal variable(s) of interest.

We consider the linear predictor

$$E^*(Y_i|W_i) = W_i'\beta$$

where E^* denotes the linear predictor such that β are the minimizer of the expected squared error loss $E((Y - E^*(Y_i|W_i))^2)$.[5] I will assume W_i includes a constant for purposes of linear regression.

Then, recall that

$$\beta = E(W_iW_i')^{-1}E(W_iY_i).$$

Note that β is a population object, defined based on two moments.

Next recall that the least-squares estimator of β is

$$\begin{aligned} b(Y_n, W_n) &= \left(n^{-1}\sum_{i=1}^{n} W_iW_i'\right)^{-1}\left(n^{-1}\sum_{i=1}^{n} W_iY_i\right) && (3)\\ &= (W_n'W_n)^{-1}W_n'Y_n. && (4) \end{aligned}$$

Recall that b is a *function* of random variables Y_n and W_n (an estimator) and as such also a random variable. Since we can't directly study $E(b)$, we study instead $E(b|W_n)$, which focuses on the conditional distribution of $Y_i|W_i$.[6]

If we consider $E(b(Y_n, W_n|W_n = w))$, we see

$$E(b(Y_n, W_n|W_n = w)) = (w'w)^{-1}w'E(Y_n|W_n = w).$$

When we are correctly specified, and the conditional expectation $E(Y_n|W_n) = W_n\beta$, then we have:

$$E(b(Y_n, W_n|W_n = w)) = \beta.$$

Since this is true for any w, we can use the law of iterated expectations and $E(E(b(Y_n, W_n|W_n = w))) = \beta$.

We will now turn to inference in this setting.

Model-based inference

Given our linear project, $E^*(Y_i|W_i) = W_i'\beta$, we write

$$Y_i = W_i'\beta + \epsilon_i,$$

where ϵ_i denotes the error term that is mechanically orthogonal to W_i. As we saw above, we'll often consider the uncertainty in the sampling *conditional* on W_i, and hence the uncertainty that drives our estimate is from ϵ_i (the unexplained part of Y_i). Restating our estimator from before,

$$\hat{\beta} = \beta + (\mathbf{W}_n'\mathbf{W}_n)^{-1}\mathbf{W}_n'\epsilon_n.$$

[5] Note that this is the traditional OLS estimator, and if we want to allow for more flexible functional forms of W_i, we'll need to include higher-order interactions and functions, etc.

[6] Why can't we study $E(b)$? The non-linearity makes it hard to study expectations – the expectation of a ratio is not equal to the ratio of expectations.

Typically we take $\mathbf{W}_n$ as given, and so the uncertainty (in the model based world) is driven by $\boldsymbol{\epsilon}_n$.

Now we consider the variance of $\hat{\beta}$. Formally, we see that this revolves around the structure of $\mathbb{E}(\boldsymbol{\epsilon}_n\boldsymbol{\epsilon}_n'|\mathbf{W}_n) = \Omega_n$:

$$\begin{aligned} \mathbb{V}(\hat{\beta}|\mathbf{W}_n) &= (\mathbf{W}_n'\mathbf{W}_n)^{-1}\mathbf{W}_n'\mathbb{E}(\boldsymbol{\epsilon}_n\boldsymbol{\epsilon}_n'|\mathbf{W}_n)\mathbf{W}_n(\mathbf{W}_n'\mathbf{W}_n)^{-1} &(5)\\ &= (\mathbf{W}_n'\mathbf{W}_n)^{-1}\mathbf{W}_n'\Omega_n\mathbf{W}_n(\mathbf{W}_n'\mathbf{W}_n)^{-1} &(6)\end{aligned}$$

Everything pivots around the structure of $\mathbb{E}(\boldsymbol{\epsilon}_n\boldsymbol{\epsilon}_n'|\mathbf{W}_n) = \Omega_n$.

So far, we have assumed that the draws are independent, so we already know that Ω_n is a diagonal matrix.[7] To simplify further, we can consider is homoskedasticity, where $\Omega = \sigma^2 I_n$. This simplifies our variance:

$$\mathbb{V}(\hat{\beta})_{homoskedastic} = \sigma^2(\mathbf{W}_n'\mathbf{W}_n)^{-1} \tag{7}$$

What is the content of this assumption? Beyond $Cov(\epsilon_i, \epsilon_j) = 0$, it assumes $Var(\epsilon_i|W_i) = Var(\epsilon_i)$.[8]

Comment 1

Consider the linear regression model when posed as potential outcomes with unobserved heterogeneity and strong ignorability:

$$\begin{aligned} Y_i &= \alpha + D_i\beta + \epsilon_i\\ \alpha &= E(Y_i(0))\\ \beta &= E(Y_i(1)) - E(Y_i(0))\\ \epsilon_i &= D_i\Bigg(\underbrace{(Y_i(1) - Y_i(0))}_{\beta_i} - \beta\Bigg) + \Bigg(Y_i(0) - E(Y_i(0))\Bigg). \end{aligned}$$

Hence, the assumptions about $Var(\epsilon_i|D_i)$ relate directly the extent of heterogeneity in the treatment effect. Namely,

$$\begin{aligned} Var(\epsilon_i|D_i = 1) &= Var(\beta_i|D_i = 1) + Var(Y_i(0)|D_i = 1) &(8)\\ Var(\epsilon_i|D_i = 0) &= Var(Y_i(0)|D_i = 0). &(9)\end{aligned}$$

These can only satisfy homoskedasticity and be equal if $Var(\beta_i) = 0$, and hence the treatment effect is constant. Under random assignment, the latter terms are equal by construction.

A feasible estimator for the homoskedastic variance estimand, where k is the number of regressors in β (excluding the constant), follows using empirical analogs:

$$\hat{\mathbb{V}}(\hat{\beta})_{homoskedastic} = \hat{\sigma}^2(\mathbf{W}_n'\mathbf{W}_n)^{-1} \qquad \hat{\sigma}^2 = (n-k-1)^{-1}\hat{\boldsymbol{\epsilon}}_n'\hat{\boldsymbol{\epsilon}}_n \tag{10}$$

If we instead want to allow $Var(\epsilon_i|W_i) = \sigma^2(W_i)$ to vary with W, this implies a potentially complicated function $\sigma^2(W_i)$. But, as

[7] Why? Verify this for yourself if this is not clear.

[8] This means that conditional on $W = w$, the variance of Y is the same, regardless of the value of w. That's pretty strong.

it turns out, the estimator for $\mathbb{V}(\hat{\beta})$ is quite straightforward. This is often referred to as the "robust" or EHW estimator [Eicker, 1963, Huber et al., 1967, White, 1980]:

$$\hat{\mathbb{V}}(\hat{\beta})_{EHW} = (\mathbf{W}_n'\mathbf{W}_n)^{-1}\sum_i \hat{\epsilon}_i^2 W_i W_i' (\mathbf{W}_n'\mathbf{W}_n)^{-1}.$$

Example 1
Consider the case where $W_i = (1, D_i)$, and D_i is a dummy treatment variable. Then, the variance of the coefficient on D_i reduces to

$$\hat{\mathbb{V}}(\hat{\beta})_{homoskedastic} = \frac{\hat{\sigma}^2}{n_0} + \frac{\hat{\sigma}^2}{n_1} = \frac{\hat{\sigma}^2}{n}, \qquad \hat{\mathbb{V}}(\hat{\beta})_{EHW} = \frac{\hat{\sigma}^2(0)}{n_0} + \frac{\hat{\sigma}^2(1)}{n_1},$$

where $n_0 = \sum_i (1 - D_i)$, $n_1 = \sum_i D_i$ and $\hat{\sigma}^2(x)$ is the estimated variance of ϵ_i for observations with $D_i = x$.

We then consider confidence intervals based around distributional assumptions. Recall that our distributional assumptions come from considering the following statistics:

$$T = \frac{\hat{\beta} - \beta}{\sqrt{\hat{\mathbb{V}}/\nu}}$$

When we assume that the distribution of ϵ is homoskedastic and Normal, we know the exact distribution of T. That's because β is consequently the sum of independent normals, and the variance of $\hat{\beta}$ is a scaled chi-squared (since a squared normal is chi-squared). This is the basis for the t-distribution.[9]

Without homoskedastic Normality, the distribution for T is only Student-t asymptotically. This approximation works pretty well in many settings, but there are edge cases where issues can arise.[10] One straightforward edge case we'll consider now is when n_1 and n_0 are not both simultaneously growing large.

Confidence intervals, finite sample performance, and the Behrens-Fisher problem

Note that in our discussion above, the feasible estimator for $\hat{\sigma}^2(x)$ was not made explicit. For the homoskedastic case, we have a simple estimator in Equation (10). We have adjusted for k in this estimator to account for the bias that arises from our estimation of β. A similar adjustment needs to occur for the heteroskedastic case to account for the estimation of the parameters as well. Note that this is a *finite sample* adjustment, since it will not matter as n gets large.

It is worth highlighting the different relevant adjustments that get made in the heteroskedastic case, which will matter for practical

[9] E.g. $T = Z/\sqrt{V/\nu}$, where $Z \sim \mathcal{N}(0,1)$, $V \sim \chi^2(\nu)$

[10] Edge cases that you likely know well include weak IV and unit roots. I call them edge cases because they usually involve sending a parameter to a boundary case, such as the first-stage coefficient to zero, or AR parameter to one.

inference.[11] The original EHW estimator is biased, and does not adjust for the k parameters. MacKinnon and White [1985] propose a second estimator, referred to as the HC2 estimator, that adjusts for the bias in the variance estimator.[12] It's exactly unbiased in the binary treatment case, but this is not generally true. In the binary case, it adjusts simply by subtracting one:

$$\hat{\sigma}^2(d) = \frac{1}{n_d - 1} \sum_i D_i (Y_i - \overline{Y}_d)^2.$$

So far, we have just discussed different estimators for $\mathbb{V}$. Recall that we want to use these to make confidence intervals, based on the distribution of T. That distribution requires an assumption about the degrees of freedom for T. Then, we can construct 95% confidence intervals based on these asymptotic results:

$$\text{CI} = \left(\beta - t^{n-2}_{0.975} \times \sqrt{\hat{\mathbb{V}}}, \beta + t^{n-2}_{0.975} \times \sqrt{\hat{\mathbb{V}}}\right)$$

where t^n_q is the qth quantile the t distribution with degrees of freedom n. The issue is that the degrees of freedom are not clear in the heteroskedastic case. Why? Because the variance is the weighted sum of *different* Chi-squared distributions. This is very concrete in Example 1, where the denominator of the scaling for the variance is driven by the relative size of the treatment and control groups.

We now consider the Behrens-Fisher problem. Imagine that $n_0 >> n_1$, that is, there are few treated units relative to the control.[13] Then, the distribution is really driven by $\sigma^2(1)/n_1$, and n_1 is the correct degrees of freedom. This makes a big difference! Contrast the degrees of freedom between the two cases: $t^3_{0.975} = 3.182$ vs. $t^{28}_{0.975} = 2.048$.[14] This naturally creates much wider confidence intervals for a given dataset, which implies that the coverage under the n_0 degrees of freedom would have much lower coverage than the n_1 degrees of freedom. We can see this coverage difference in simulations done by Imbens and Kolesar [2016] in Figure 1.

The estimator with the best performance in this setting is the Bell-Maccaffrey adjustment, which is a generalization of the HC2 estimator. This estimator adjusts for the degrees of freedom issue by finding the parameter K which creates a t-distribution that most closely matches the first two moments of the dispersion of the estimated $\hat{V}_{HC2}$ around the estimand.[15] In the binary case, this reduces to

$$K_{BM} = \frac{(n_0 + n_1)^2 (n_0 - 1)(n_1 - 1)}{n_1^2 (n_1 - 1) + n_0^2 (n_0 - 1)}.$$

Some intuition arises in the case when n_0 and n_1 are similar: we get $K_{BM} = n - 2$, which is the degrees of freedom in the homoskedastic case. If $n_0 >> n_1$, then we get $K_{BM} \approx n_1$. This is a very intuitive

[11] See the discussion in Comment 2 for implementations, and the implications of using biased estimators in Figure 1.

[12] There is also a third estimator, referred to as HC3, from MacKinnon [2012], that also adjusts for the bias in the variance estimator. It is quite conservative (it's biased upwards in the case of binary treatment). For the binary case, it is given by

$$\hat{\mathbb{V}}(\hat{\beta})_{EHW} = \hat{\sigma}^2(0)\frac{n_0}{(n_0 - 1)^2} + \hat{\sigma}^2(1)\frac{n_1}{(n_1 - 1)^2}.$$

I provide it for completeness, but it is not widely used.

[13] These results and discussion come from Imbens and Kolesar [2016].

[14] Notably, this issue starts to disappear as the minimum size gets large.

[15] This is done under the assumption of homoskedasticity, for *just* the purposes of estimating the degrees of freedom.

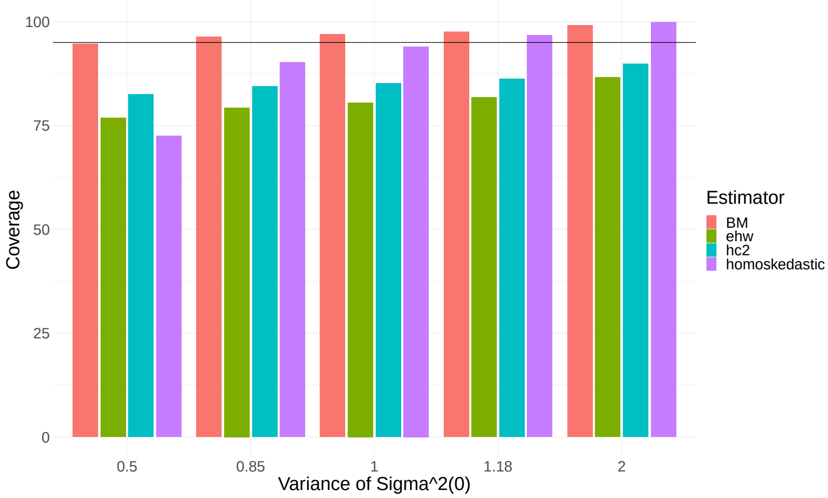


Figure 1: These simulation results come from Imbens and Kolesar [2016] and study the impact of unbalanced treatment and control. In the simulation, $n = 30, n_0 = 27, n_1 = 3$. The coverage of the different estimators discussed in **??** are shown. The Bell-Maccaffrey adjustments are the most accurate, and EHW does poorly in all cases. The data is simulated to be Normally distributed conditional on treatment, and the Variance of the error term is allowed to vary.

result, and lines up with what we'd want. Imbens and Kolesar [2016] recommend this estimator to researchers in practice.

The Behrens-Fisher issue (and Bell-Maccaffrey solution) generalizes to general regression setting, even when the treatment is not binary. The key idea is that the variance we're scaling by is not a Chi-squared with the full degrees of freedom n. The approximation that we use matches the degrees of freedom to get the first and second moment as close as possible to the "right" chi-squared.[16]

Comment 2
What are the statistical packages that we use in these cases, and what's doable?

- *Stata uses the EHW standard errors by default (with the* `robust` *command). But, I believe as of Stata 17, it does the correct finite sample adjustment for the degrees of freedom to make it unbiased in the simple case. You can use* `vce(hc2)` *to get the HC2 estimator, and* `vce(hc3)` *to get the HC3 estimator. For the Bell-Maccaffrey adjustments, see* `reg_sandwich`*.*
- *In R, there are many pacakges, but the ones I would look see* `estimatR`*,* `clubSandwich`*, and Kolesar's github repo.*

[16] These issues can rear their head when the the regressor of interest is highly skewed (e.g. a log-normal right hand side variable). The intuition comes from the fact that the distribution of the regression will affect the distribution of $(W_n' W_n)^{-1} W_n' \epsilon_n$, warping the finite sample behavior of the sum of the ϵ_n sum.

Example 2
You might be asking yourself, why do I have to care about these things? For a lovely discussion of finite sample issues, you should peruse Data Colada's discussion of Alwyn Young's QJE paper on randomization inference. The paper is a great example of how finite sample issues can matter in practice.

To quote the post: "In this post I show that this conclusion only holds when relying on an unfortunate default setting in Stata. In contrast, when regression results are computed using the default setting in R [1], a setting that's also available in Stata, a setting shown over 20 years ago to be more appropriate than the one used by Stata... the supposed superiority of the randomization test goes away."

Bootstrap One nice alternative approach to distributional assumptions on T is to use the bootstrap. The bootstrap is a general purpose tool for constructing confidence intervals, and it can be used in the context of linear regression. The idea is to resample the data with replacement, and then re-estimate the model. This comes in many forms, but the most intuitive and straightforward form is called the non-parametric bootstrap, which would involve resampling n observations of (Y_i, W_i) from the data with replacement, and then re-estimating the model. After B samples are re-estimated, this will give us a distribution of the parameter and we can describe the statistical properties of this distribution (e.g. the 95% interval of this distribution). It is also plausible to construct t-statistics $t_b = (\hat{\beta}_b - \beta^0)/\sqrt{V_b}$ for each bootstrap sample b and null hypothesis β^0, and use this instead of $\hat{\beta}$. Since the t-statistic is asymptotically pivotal, it can have nice properties.

However, the non-parametric bootstrap can have issues if the sample is small or the regressors are skewed [Imbens and Kolesar, 2016], as the additional noise introduced by resampling creates worse distributional approximations. One very successful bootstrap alternative is the wild bootstrap. Concretely, this works as follows (see Davidson and Flachaire [2008] for details):

1. Estimate the linear model $\hat{\beta}$ and obtain residuals $\hat{\epsilon}_i$.

2. In each bootstrap step b:

   (a) For each observation i, the X_i is fixed, along with $\hat{\beta}$ and $\hat{\epsilon}_i$. We then draw a binary variable $U_{i,b}$ that is either 1 or -1 with equal probability. We set $Y_{i,b} = \hat{\beta}X_i + U_{i,b}\hat{\epsilon}_i$.

   (b) With the new dataset, we re-estimate the model and construct a t-statistic $t_b^1 = (\hat{\beta}_b - \hat{\beta})/\sqrt{\hat{V}_b}$, where $\sqrt{\hat{V}_b}$ is the standard error.

3. With the full set of t_b, we can construct a confidence interval by calculating the $q_{0.95}(|t_b|)$, the 0.95 quantile of $|t_b|$, and using it in the place of our usual critical value:

$$\text{CI}_{WILD} = \left(\hat{\beta} - q_{0.95}(|t_b|)\sqrt{\hat{V}}, \hat{\beta} + q_{0.95}(|t_b|)\sqrt{\hat{V}}\right)$$

Combining Sampling- and Design-based uncertainty

How should we be thinking about inference anyway? What's the error in ϵ *mean*? The thought experiment typically comes from a sampling perspective – we consider that this is a small sample from a broader population, and uncertainty comes from whether the estimates reflect the true underlying population. Note that this contrasts with our design-based thought experiment!

This starts to get very confusing when thinking about some settings. For example, how do we think about sampling "new states" when we have all 50 states? Worse yet, what if we have access to all the census data? We observe the whole population! What's the uncertainty in our estimates then? In estimation of causal effects, we still uncertainty because there's uncertainty driven by the fundamental problem of causal inference!

Using work from Abadie et al. [2020], we will now consider two sources of uncertainty: sampling and design. There exists a population of size N and a sample of size $n \leq N$.[17] Let $R_i = \{0, 1\}$ denotes whether or not an observation is in the sample. There are also potential outcomes $Y^*(D_i)$. Now we have both sampling uncertainty (e.g. does our sample reflect the population) and design uncertainty (e.g. does the causal comparison reflect the true causal effect). We can now combine the two sources of uncertainty to get a better understanding of the variance of our estimator.

I will focus on just binary case of a single treatment, but the paper considers full regression setting. We consider three estimands:

1. $\theta^{descr} = N_1^{-1} \sum_{i=1}^N D_i Y_i - N_0^{-1} \sum_{i=1}^N (1 - D_i) Y_i$
2. $\theta^{causal,sample} = n^{-1} \sum_{i=1}^N R_i (Y_i^*(1) - Y_i^*(0))$
3. $\theta^{causal} = N^{-1} \sum_{i=1}^N (Y_i^*(1) - Y_i^*(0))$

We have a single estimator we can consider:

$$\hat{\theta} = n_1^{-1} \sum_{i=1}^N R_i D_i Y_i - n_0^{-1} \sum_{i=1}^N R_i (1 - D_i) Y_i$$

The key point of paper – the variance of this estimator depends on what we condition on. If we condition on D, we focus on sampling

[17] My n, N are reversed from the paper.

uncertainty. If we condition on R we focus on causal uncertainty within sample. If we condition on neither, we focus on causal uncertainty as well as sampling uncertainty.

What are these variance estimands? Let S_x denote the population variance for each potential outcome, and S_θ denote the population variance of the treatment effect outcomes (which recall, we cannot directly estimate).[18] Then, we have the following variance estimands:

1. Sampling: $E(Var(\hat{\theta}|\mathbf{D}, n_1, n_0)|n_1, n_0) = \frac{S_1^2}{n_1}\left(1 - \frac{n_1}{N_1}\right) + \frac{S_0^2}{n_0}\left(1 - \frac{n_0}{N_0}\right)$,
2. Design: $E(Var(\hat{\theta}|\mathbf{R}, n_1, n_0)|n_1, n_0) = \frac{S_1^2}{n_1} + \frac{S_0^2}{n_0} - \frac{S_\theta^2}{n_1+n_0}$
3. Both: $Var(\hat{\theta}|n_1, n_0) = \frac{S_1^2}{n_1} + \frac{S_0^2}{n_0} - \frac{S_\theta^2}{N_1+N_0}$

The S_θ^2 term is what we usually ignore (because not feasibly estimable). Consider the following thought experiments: let n get close to N, or let n get small relative to N. This will move around the sampling and total variance estimands, but not the sample causal estimand. Next, notice that the difference between Sampling and Design is not obvious. Sampling can be small if n is close to N, but when S_θ^2 is large, that can make the design variance small.

Clustering and generalizing Ω

This ignored any sort of unusual correlation structure in Ω, and assumed random assignment. In many cases, we don't have that. Instead, Ω has a clustering structure. This can get quite complex. Let's start with the simple case of known clusters. E.g. units are people, and clusters are cities, counties or states. For today, we're ignoring the very important question of panel data. We'll come back to that later.

Let C_i denote unit i's cluster assignment. A very simple version of Ω is now

$$\Omega_{ij} = \begin{cases} \sigma^2 & \text{if} \quad i = j \\ \rho\sigma^2 & \text{if} \quad C_i = C_j \;\&\; i \neq j \\ 0 & \text{if} \quad C_i \neq C_j \;\&\; i \neq j \end{cases} \tag{11}$$

This matrix can also be more unstructured. For exmaple, one could have a flexible block diagonal with $\Omega_{ij} = \sigma_{ij}$ if $C_i = C_j$. A key issue that arises here is in the relative size of the block versus the number of blocks. See Hansen [2007] for a discussion of one example.

Let's start with a model-based approach to this. Denote the number of clusters by K. Following Liang and Zeger [1986], the estimator

[18] This variance is $S_\theta = \frac{1}{N-1}\sum_i \left(Y_i(1) - Y_i(0) - N^{-1}\sum_i(Y_i(1) - Y_i(0))\right)^2$.

for the variance of $\hat{\beta}$ is

$$\hat{\mathbb{V}}(\hat{\beta}|\mathbf{W}_n, \mathbf{C}_n)_{LZ} = (\mathbf{W}_n'\mathbf{W}_n)^{-1}\left(\sum_{k=1}^{K}\mathbf{W}_{k,n}'\hat{\boldsymbol{\epsilon}}_{k,n}\hat{\boldsymbol{\epsilon}}_{k,n}'\mathbf{W}_{k,n}\right)(\mathbf{W}_n'\mathbf{W}_n)^{-1}. \tag{12}$$

This estimator allows for flexible covariance within the block, and assumes that the size of the block is fixed, and that K is large.[19] Historically, clustering in this setting has focused the structure of Ω. Why? Well, take the simple case in Equation (11). In this case,

$$\mathbb{V}(\hat{\beta}) = \mathbb{V}_{homoskedastic} \times \left(1 + \rho_\epsilon \rho_W \frac{n}{K_n}\right), \tag{13}$$

where ρ_ϵ and ρ_W are the within-cluster correlations of each r.v. This makes you think that these are the main terms that matter, and more generally it's about getting the structure of Ω right. E.g., better to err on the conservative side and let the blocks be large.

However this intuition is *not* correct in contexts with any meaningful heteroskedasticity. Abadie et al. [2023] can generate an example with tiny within-cluster correlation, and large clusters (with many clusters) where the Liang-Zeger estimator $\hat{\mathbb{V}}(\hat{\beta})_{LZ}$ is large and $\hat{\mathbb{V}}(\hat{\beta})_{EHW}$ is small. How come? Recall from Comment 1 that the variance of the error term depends on two pieces: the variance of the poetntial outcome, and the variance of the tratement effect, as it correlates with treatment. Namely, it's all about the correlation *between* W and ϵ, and heterogeneity in our effects across clusters.

In Abadie et al. [2023], they construct an example where $N = 10,000,000$ with 100 equal sized clusters. There is a binary W with *equal* probability. THere are significant heterogeneous effects of W across clusters – some clusters have positive effect, some have negative. Overall ATE is 0. What does that mean intuitively? If there is heterogeneity in effects, it causes correlation between treatment and residual.

So why do the two standard error estimators vary so much? To quote the Abadie et al. [2023] working paper:

> The reason for the difference between the EHW and LZ standard errors is simple, but reflects the fundamental source of confusion in this literature. Given the random assignment both standard errors are correct, but for different estimands. The LZ standard errors are based on the presumption that there are clusters in the population of interest beyond the 100 clusters that are seen in the sample. The EHW standard errors assume the sample is drawn randomly from the population of interest. It is this presumption underlying the LZ standard errors of existence of clusters that are not observed in the sample, but that are part of the population of interest, that is critical, and often implicit, in the model-based motivation for clustering the standard errors. It is of course explicit in the sampling design literature (e.g., Kish [1965]). If we changed the set up to one where the population of 10,000,000 consisted of say 1,000 clusters, with 100 clusters drawn at random, and then sampling units randomly from those sampled clusters, the LZ standard errors would be correct, and the EHW standard errors would be incorrect. Obviously one cannot tell from the sample itself whether there exist such clusters that are part of the population of interest that are not in the sample, and therefore one needs to choose between the two standard errors on the basis of substantive knowledge of the study design.

[19] In fact, the degrees of freedom are defined by the size of K.

What are the key takeaways from this paper? First, cluster your regression at the unit of randomization. Being conservative can be quite bad! It depends on what you are trying to do. The traditional advice of being as conservative as necessary is likely misguided. Fixed effects do NOT remove the need for clustering. We'll revisit this in panel settings.

Discussion Questions 1
What is the "unit of randomization" in a case like CARD and KRUEGER [1994]?

Comment 3 (Spatial and Network Error)
Things get more complicated with more general error structures. Consider two additional cases:

- *Spatial correlation =* $\rho_{ij} = f(d_{ij})$*, where* d_{ij} *is a function of some economic distance.*
- *Social network correlation =* $\rho_{ij} = f(d_{ij})$*, where* d_{ij} *is a function of path length in a network*

This can matter especially when SUTVA is violated However, Barrios et al. (2012) show that, under SUTVA, if treatments are randomly assigned at a given cluster level, we can ignore the broader spatial correlations
Conley (1999) provides a flexible way to consider clustering on spatial distances. Consider our matrix Ω *again. Now,* Ω_{ij} *is a function of the distance,* d_{ij}*, between each person. Unfortunately, this means that every person can be correlated. Key assumption – the correlation declines with distance. Hence, far away distances matter less in practice. Hence, when we estimate this, we "window" our estimator (this is exactly the same as Newey-West estimators). Then we allow correlation as in the Liang-Zeger estimator, as a function of distances. This estimator is consistent for general forms of spatial correlation*

- *Estiators available in both Stata and R*

Example 3
Consequences of ignoring spatial correlation
Spatial correlation can be a big deal. Consider the analogy to time series.

- *A big rule: worry about highly autocorrelated data! Can inflate your t-statistics substantially*
- *Why? Because if we treat observations as independent, we will infer more information than actually exists*

Kelly (2019) claims that spatial correlation in outcomes can cause this same issue. Consider a regression of some modern outcome, e.g. city income, on a historical characteristic, such as colonial boundaries

- *Claim in Kelly (2019) is that t-statistics in these types of regressions are grossly amplified by spatial correlation*
- *Fixable with Conley standard errors?*
- *This is a huge deal for a lot of literatures (economic history especially) – matters for corporate governance literature too (LLSV)*

Concluding thoughts

This stuff is *hard*. We are doing the simplest case (linear regression) and still have lots of questions. As always, asking what the knowable estimand is can be very helpful. Next, if you are unsure, it is very useful to consider simulating data. [20] In many cases, there is not an obvious "best" answer, and simulating your data is the best solution. This is because many results are asymptotic in nature, and hence approximations.

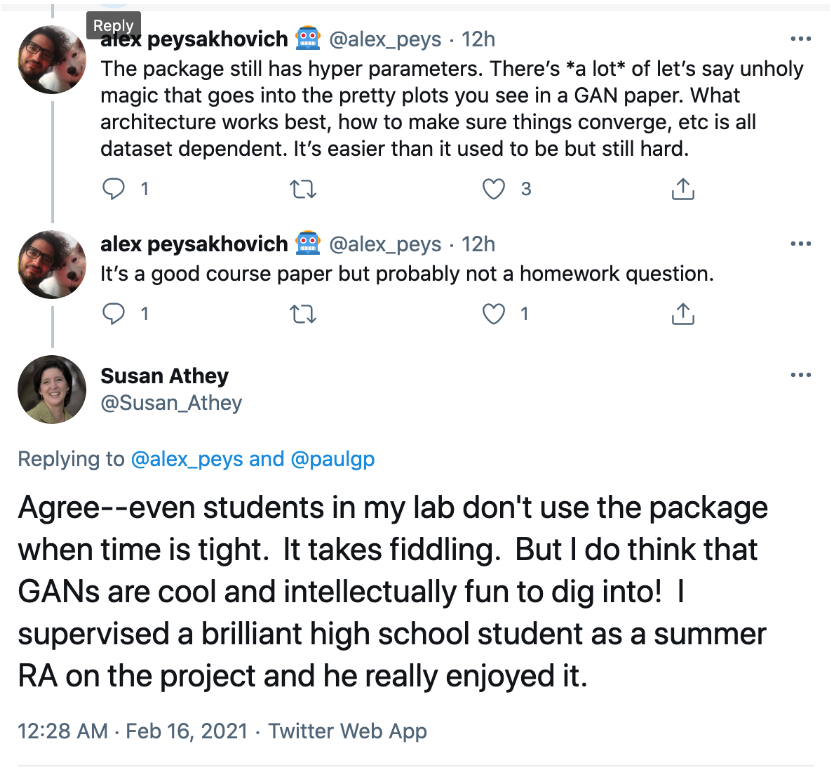
So how does one implement a simulation? Goal is to generate data that matches your dataset's distributions. However, for very simple simulations, you'll have to make parametric assumptions that may not match your actual data. Athey et al. (2020) propose a method for matching the data as closely as possible, using a Generative Adverserial Network. In other words, construct distributions that match the "true" data as closely as possible

- Computationally expensive, but great way to evaluate performance
- Code is available here: https://github.com/gsbDBI/ds-wgan
- Docs are here: https://github.com/gsbDBI/ds-wgan

However this stuff is really hard to implement. If you intuitively

[20] This is the approach advocated in Blair et al. [2023].

know the issue, try doing something simple with normals Or try bootstrapping!



alex peysakhovich 🤖 @alex_peys · 12h

The package still has hyper parameters. There's *a lot* of let's say unholy magic that goes into the pretty plots you see in a GAN paper. What architecture works best, how to make sure things converge, etc is all dataset dependent. It's easier than it used to be but still hard.

alex peysakhovich 🤖 @alex_peys · 12h

It's a good course paper but probably not a homework question.

Susan Athey
@Susan_Athey

Replying to @alex_peys and @paulgp

Agree--even students in my lab don't use the package when time is tight. It takes fiddling. But I do think that GANs are cool and intellectually fun to dig into! I supervised a brilliant high school student as a summer RA on the project and he really enjoyed it.

12:28 AM · Feb 16, 2021 · Twitter Web App

References

Alberto Abadie, Susan Athey, Guido W Imbens, and Jeffrey M Wooldridge. Sampling-based versus design-based uncertainty in regression analysis. *Econometrica*, 88(1):265–296, 2020.

Alberto Abadie, Susan Athey, Guido W Imbens, and Jeffrey M Wooldridge. When should you adjust standard errors for clustering? *The Quarterly Journal of Economics*, 138(1):1–35, 2023.

Graeme Blair, Alexander Coppock, and Macartan Humphreys. *Research Design in the Social Sciences: Declaration, Diagnosis, and Redesign*. Princeton University Press, 2023.

DAVID CARD and ALAN B KRUEGER. Minimum wages and employment: A case study of the fast-food industry in new jersey and pennsylvania. *The American Economic Review*, 84(4):772–793, 1994.

Russell Davidson and Emmanuel Flachaire. The wild bootstrap, tamed at last. *Journal of Econometrics*, 146(1):162–169, 2008. ISSN

0304-4076. DOI: https://doi.org/10.1016/j.jeconom.2008.08.003. URL https://www.sciencedirect.com/science/article/pii/S0304407608000833.

Friedhelm Eicker. Asymptotic normality and consistency of the least squares estimators for families of linear regressions. *The annals of mathematical statistics*, 34(2):447–456, 1963.

Christian B Hansen. Asymptotic properties of a robust variance matrix estimator for panel data when t is large. *Journal of Econometrics*, 141(2):597–620, 2007.

Peter J Huber et al. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 221–233. Berkeley, CA: University of California Press, 1967.

Guido W Imbens and Michal Kolesar. Robust standard errors in small samples: Some practical advice. *Review of Economics and Statistics*, 98(4):701–712, 2016.

Kung-Yee Liang and Scott L Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986.

James G MacKinnon. Thirty years of heteroskedasticity-robust inference. In *Recent advances and future directions in causality, prediction, and specification analysis: Essays in honor of Halbert L. White Jr*, pages 437–461. Springer, 2012.

James G MacKinnon and Halbert White. Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of econometrics*, 29(3):305–325, 1985.

Halbert White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: journal of the Econometric Society*, pages 817–838, 1980.

Lecture 6 – Linear Regression 2 — Semiparametrics and Visualization

Paul Goldsmith-Pinkham

February 1, 2024

Linear Regression: Why so Popular

Linear regression is incredibly popular as a tool. Why? Many reasons:

- Fast (easy analytic solution and matrix inversion has gotten better)
- Efficient (under some settings, OLS is BLUE)

My view: linear regressions is

1. an intuitive summary of data relationships
2. A good default – many "better" options are only good in some settings, and linear regression is not bad in many
3. Does a good job with many of the things we throw at our models (high dimensional fixed effects, lots of data)

Today: how to stay in the world of linear regression as much as possible, improving our presentation

As a side goal, we will do a discussion on good visualization practice

Lecture 9 - Discrete Choice and GLM

Paul Goldsmith-Pinkham

February 28, 2024

We are now going to generalize our estimation problem beyond linear models like linear (and quantile) regression, and consider more complex objective functions. This will initially be motivated by the binary choice model, but will be more generally applicable to a wide range of problems. This will lead to us covering a wide range of topics, including binary choice models, generalized linear models (GLMs), numerical estimation methods for non-linear models, inconsistency of non-linear models with many parameters, and the challenges of estimating models with multiple discrete choices.

Conceptually, we will be considering minimizing *objective functions* as a general case of minimizing *squares*.

Binary choice

Consider the following binary outcome problem: let Y_i denote if person i is a homeowner, and X_i includes three covariates: income, age and age^2 (plus a constant). How should we model the relationship between X and Y? Conceptually, a very general form would consider

$$Y_i = F_i(X_i),$$

where F_i could vary by individual. However, this doesn't seem like a very good model for considering estimands, such as "how much does homeownership increase with a 10k increase in income?"[1] In many ways, this is similar to the questions related to binscatter and other semiparametric models.

The potentially issues with blithly assuming a linear model for $F_i(X_i)$ becomes very apparent in the context of a binary dependent variable. Say we model this outcome using a linear regression (this is often called a linear probability model), assuming strong ignorability or just $E(\epsilon_i|X_i) = 0$:

$$E(Y_i|X_i) = Pr(Y_i = 1|X_i) = X_i\beta \qquad \rightarrow Y_i = X_i\beta + \epsilon_i \tag{1}$$

The problems with modeling Y in this way is twofold. First, since the outcome is binary, the error structure will be bimodal and unusual looking. To see this, consider $\varepsilon_i = Y_i - X_i\beta$, and consider how ε_i changes for $Y_i = 0$ vs 1. For a given X_i, it is exactly bimodal (like the outcome). One implication of this is that $V(Y|X) = X_i\beta(1 - X_i\beta)$, and you'll have pretty significant heteroskedasticity. This is solveable

[1] Formally, this would look something like $E(dF_i(X_i)/dX_i|X_i)$, and we would need to make some assumptions on F_i to make progress. That's what we'll do now.

using robust standard errors, but does mean that a normal approximation with the error is a poor one.

Seoncd, except under some special circumstances, it's very likely that the predicted values of Y_i will be outside of $[0,1]$. What's an example where they will not be? Discrete exhaustive regressors! Why? Discrete exhaustive regressors are the one setting where you can guarantee that the model is correctly specified. When the model is misspecified, it is quite possible that the model will extrapolate in a way such that there will be values outside support.

How does this impact our causal estimates? If the model is correctly specified, we can generate counterfactual predictions of the outcome. If not, then we get a linear approximation that may be nonsensical.

Table 1: LPM model estimates

variable	linear est.	std.error
Intercept	0.0242	0.0410
age	0.0220	0.0017
age^2	-0.0002	0.0000
income /10k	0.0069	0.0007

Example 1 (LPM estimates of homeownership)
We estimate the linear model in Table 1. and note tthat if income were strictly ignorable, we could say that 10k increase in income leads to 0.69 p.p. increase in the probability of homeownership. But, the predicted probability of homeownership would range from 0.283 *to* 1.78. *Oops.*

Modeling discrete choice

There are two ways to think about how we think about this estimation problem. These are *not* mutually exclusive, and it is important to note that both of these approaches are very focused on the *model-based* aspect of estimating causal effects.

The first is a statistical view. How can we model the statistical process for Y_i better? In other words, can we fit the outcome model better? Consider $X_i\beta$ as the conditional mean of some process, what's the statistical model that fits with this? This is a case of what's termed "Generalized Linear Models" (GLM)

A second way to view this is as an structural (economic) choice problem. Most models of binary outcome variabels assume a latent index, on the utility of choosing Y_i:[2]

$$Y_i^* = X_i\beta + \varepsilon_i, \qquad Y_i = \begin{cases} 1 & Y_i^* > 0 \\ 0 & Y_i^* \leq 0. \end{cases} \tag{2}$$

As we will now see, both approaches do arrive at a similar modeling conclusion, but the latter model will naturally accomodate choices.

A natural approach in either of these is to make a distributional assumption about ε_i. Two common assumptions:

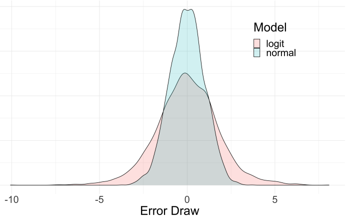
[2] The careful reader will note that analogy to the Heckman model on treatment choice.

1. ε_i is conditionally normally distributed (probit), such that $Pr(Y_i = 1|X_i) = \Phi(X_i\beta)$
2. ε_i is conditionally extreme value (logistic) such that $Pr(Y_i = 1|X_i) = \frac{\exp(X_i\beta)}{1+\exp(X_i\beta)}$

Note that these are not, in the binary setting, deeply substantive assumptions. In Figure 1, we see that there are very minor differences in the thickness of the tails for a logit vs. normal error, but they're both symmetric and centered around zero.[3] One downside for probit models is that there's no closed form solution for Φ, the CDF for the normal distribution:

$$\Phi(X_i\beta) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{X_i\beta} e^{-t^2/2} dt \tag{3}$$

Figure 1: Logit vs. Probit error terms



We will discuss later how to estimate β given these assumptions, but they will involve numerical optimization, as there is no closed form for β like in linear regression.

Example 1 (continued)

Consider now the same homeowner problem from Example 1, but estimated with logit. The β coefficients in Column 1 of Table 2 are hard to interpret. To see why, consider the derviative of the probability with respect to X_i:

$$\frac{\partial Pr(Y_i = 1|X_i)}{\partial X_i} = \beta\phi(X_i\beta) \qquad \textit{(Probit)}$$

$$\frac{\partial Pr(Y_i = 1|X_i)}{\partial X_i} = \beta\frac{\exp(X_i\beta)}{(1+\exp(X_i\beta))^2} \qquad \textit{(Logit).}$$

In both cases, the effect of X_i changes, depending on the value of X_i. This is a problem for interpretation. The average derivative in Column 3 is a way to get around this, but it's not a perfect solution:

$$n^{-1}\sum_i \frac{\partial E(Y|X)}{\partial X} = n^{-1}\sum_i \beta\frac{\exp(X_i\beta)}{(1+\exp(X_i\beta))^2}$$

This will calculate the derivative for every value in the sample, and then average them. This is a way to get a sense of the average effect of X_i on Y_i. We see a much larger effect of income on homeownership in the logit model than in the linear model (Column 2).

Table 2: Homeownership problem estimated with logit

	(1)	(2)	(3)
term	logit est.	linear est.	avg. deriv.
constant	-2.14	0.0242	-0.392
age	0.0903	0.022	0.0166
age^2	-0.0006	-0.0002	-0.0001
income/10k	0.0716	0.0069	0.0131

[3] Important caveat: these models only identify β up to scale. Why? The "true" model of ϵ could have variance σ^2 that is unknown. Consider if $F(X_i\beta) = \Phi(X_i\beta)$. If this were a general normal (rather than standardized with variance 1), we could just scale up the coefficients proportionate to σ and the realized binary outcome would identical. Hence, we normalize $\sigma = 1$ in most cases. This is *not* a meaningful assumption.

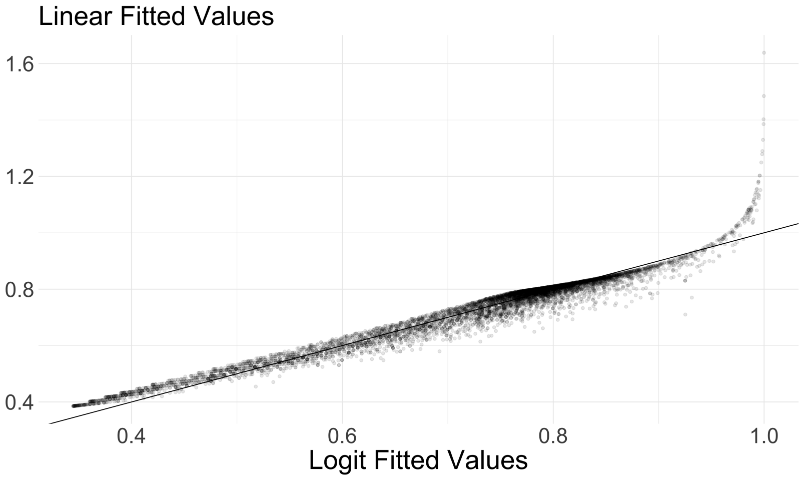


Figure 2: Linear vs. Logit model predictions

Example 1 (continued)
Figure 2 shows the predicted values of homeownership from the linear and logit models. The linear model is predicting values outside of the support of the outcome, and the logit model is not. This is one benefit of correctly specifying the model.

Generalized Linear Models (GLM)

We can generalize the intuition above, where we let the underlying distribution of ϵ be non-normal, and parameterize the mean of the distribution to be a function of $X_i\beta$. This is the idea behind Generalized Linear Models (GLM), originally formulated in Nelder and Wedderburn [1972].[4]

The overall setup of GLMs in broad strokes is to consider estimation of a **linear model** $X\beta$, which is linked to the conditional mean $E(Y|X)$ by **a link function** g: $E(Y|X) = g^{-1}(X\beta)$. The crucial underlying machinery is based on the idea that Y, the outcome, is distributed by some member of the **exponential family** of distributions. This includes the normal, binomial, Poisson, and gamma distributions, among others. Quite notably, though, as we discuss below, this distribution does not have to be correctly specified for the parameter estimates for the conditional mean to be consistent.

Some simple examples of GLMs include:

1. Logit, with a link function $g^{-1}(X_i\beta) = \frac{\exp(X_i\beta)}{1+\exp(X_i\beta)}$
2. Normal, with an identity link function $g^{-1}(X_i\beta) = X_i\beta$
3. Poisson, with a log link function $g^{-1}(X_i\beta) = \exp(X_i\beta)$

[4] Interestingly, this is very common in non-economics fields, but much less common in economics.

In essence, we can enforce a linear functional form to the *mean*, and allow the error distribution to fit the form of the data.[5]

We will now discuss the Poisson regression case in more detail, as it tends to be underused in economics, and is a very important use case. A key takeaway in GLM, like with OLS, is that it is possible to correctly specify just the conditional mean function and then robustly estimation standard errors on parameters of that function, rather than fully specifying the distribution correctly.

Poisson Regression for non-negative outcomes

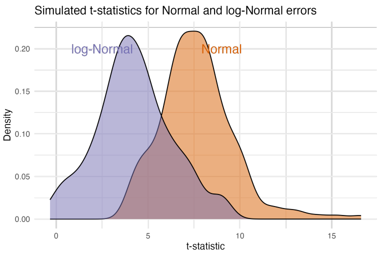
Consider an non-negative outcome $Y \geq 0$. There are a huge host of outcomes in economics and finance that are restricted to this support: investment, assets, wages, patent citations, output, and so on. We are often interested in the estimand of the partial effect $dE(Y|X)/dX$. If we estimate this conditional with linear regression (e.g. by assuming $Y_i = X_i\beta + \epsilon_i$), what are potential issues?

Mechanically, the error terms $\hat{\epsilon}_i = Y_i - X_i\hat{\beta}$ will be skewed, since Y_i is skewed. This is not on its own a huge issue, but it does suggest that the asymptotic approximation for $\hat{\beta}$ will be worse for a given n. This leads to highly influential outliers for OLS as well.[6]

Comment 1

Consider two outcomes, Y_1 and Y_2. In both cases, the true model is linear (with coefficient of 1) with respect to X, but the error term is Normal with mean zero and variance 1 in Y_1, and is log-Normal with mean zero and variance 1 in Y_2. If we simulate and estimate this model using linear regression, plotting the t-statistic of the coefficient on X for each model, we find much higher power for the model with Normal errors, rather than log-Normal errors. This reflects the lack of efficiency of OLS in the presence of non-Normal errors (but not a lack of consistency!). See Figure 3 for a visual representation of this.

Figure 3: Non-normal errors in linear regression



What are solutions to this issue? One commonly used approach is to estimate linear regressions on $\log(Y)$ instead of Y_i. This solves many of the outlier and skew issues,[7] but creates its own problems.

First, the parameters have a different interpretation. Note that the units for the outcome are different (log points). Often, these are interpreted in percentage points, since log differences are approximately equal to percentage changes.[8] This is useful, but can at times be confusing (e.g. what is the actual level effect? Sometimes a percentage effect can exaggerate or minimize a large level effect).

Second, what if $Y = 0$? This is a problem, as $\log(0)$ is undefined.

[5] It's interesting to note the underlying machinery of GLMs is similar to many of the selection and discrete choice models we've discussed and discuss today. The linear index provides an extremely convenient parameterization of the mean, but also makes some particular assumptions about the substitutibilty of the covariates.

[6] Note that one solution to this issue is to consider quantile regression instead!

[7] In the ideal case, note that a log-Normal outcome will be exactly linear after using logs.

[8] Recall $\log(Y_1) - \log(Y_0) = \log(Y_1/Y_0) = \log(1 + \Delta Y/Y_0) \approx \Delta Y/Y_0$ for $\Delta Y/Y_0$ small.

One common solution is to use $\log(1+Y)$ or $arcsinh(Y)$ [Manning and Mullahy, 2001, Ravallion, 2017, Bellemare and Wichman, 2020], which solves the second problem, but makes the first problem even worse! [Bellemare and Wichman, 2020, Aihounton and Henningsen, 2021, Cohn et al., 2022] Why these solutions? For one, they're both well-defined at $Y = 0$. Second, it has "similar" properties to the taking a log. Effectively, since the distance between $\log(1+Y)$ and $\log(Y)$ was small as Y gets large, the hope is that the differences would "wash out." It turns out, thanks to work by Chen and Roth [2023], that neither of these solutions are a good idea and that these differences do not wash out.

The key point of Chen and Roth [2023] is that percentage effects are not well-defined for outcomes that are potentially zero-valued. That is in some ways obvious – there is no way to talk about the percent increase for something where the base-level is zero. Dividing by zero is infinite! But recall that part of the goal of using log outcomes was to approximate percentage changes in the outcome due to treatments. The main result of Chen and Roth [2023] shows that for *any other function* approximating log, but defined at zero, the results will be arbitrarily sensitive to changes in units (e.g. dollars to yuan).[9]

What drives this effect? Effects close to zero and at zero. Most importantly, the *extensive margin* of moving from zero to non-zero has huge, and arbitrary, impacts on estimates on these types of rescaling. Put precisely, if you change the units of the outcome by a (e.g. $a = 100$, converting from cents to dollars), then the estimated effect will change by $\log(a)$ multiplied by the *extensive margin* effect. Note that this is fails scale equivariance, which is the property of OLS and quantile regression that usually makes a good estimator.[10]

This can have some really serious implications. Chen and Roth [2023] find that for half the papers they surveyed in the AER, the estimated effects would change by more than 100% if the units of the outcome were changed by 100 (e.g. dollars to Yen). This is a non-trivial effect!

The takeaway I want you to have: you should not be running a regression with $\log(1+Y)$ or $arcsinh(Y)$ on the left-hand side![11] So what should you do if you have a zero in your left-hand side variable? Chen and Roth [2023] suggest other ways of considering these situations:[12]

1. First, if you really need something interpretable as as a percentage effect (e.g. rescaling an ATE into percentage), you could estimate $\tau = E(Y_i(1) - Y_i(0))/E(Y_i(0))$, which scales the ATE by the baseline average. *This is the estimand targeted by Poisson regression.* There are also other normalizations one could consider. Instead of normalizing by $E(Y_i(0))$, if there is a pre-treamtent characteristic

[9] This includes both $\log(1+y)$ and $arcsinh(y)$.

[10] Note also the practical implication: if there is a big extensive margin effect, a large a has a big effect. In contrast, with a small a, then most effects will be close to zero (since they are extensive margin, and hence close to zero by definition).

[11] I have done this, historically, in my own work – we're all flawed creatures trying to inch towards better methodological implementations!

[12] These solutions are not perfect, but are motivated by a "trilemma" they prove: it is not possible to have an estimator that is simultaneously (1) an average of individual level treatment effects (2) invariant to rescaling of units and (3) point-identified without more assumptions about the joint distribution of the potential outcomes (beyond what we usually do in regression).

that is exogeneous, you could normalize by $E(Y_i(0)|W_i)$, e.g. the predicted baseline value given characteristic W_i. This captures richer heterogeneity in the baseline characteristic, and may do a better job of reducing skewness.

2. Second, you could redefine the outcome in terms of functionals of the distribution, e.g. $\tilde{Y} = F_{Y^*}(Y)$. A prominent example is looking at the rank of an individual relative to the overall individual, as in Chetty et al. [2014].
3. If the goal is to consider trade-offs in some like concave preferences, then it is plausible to specify exactly the 'value' of a person at $Y = 0$, relative to positive Y, and then explicitly evaluate the parameter that way. This has the problem of losing scale-invariance, but at least the research is explicit about how they value these issues.
4. Finally, it is plausible to directly estimate the extensive and intensive effects separately. However, the intensive effect is only partially identified; we will explore this further in later lecutres.

See Table 3 for a full set of alternative estimators.

Description	**Parameter**	**Pros/Cons**
Normalized ATE	$E(Y(1) - E(Y(0))$	Pro: Percent interpretation Con: Does not capture decreasing returns
Normalized outcome	$E(Y(1)/X - Y(0)/X)$	Pro: Per-unit-X interpretation Con: Need to find sensible X
Explicit tradeoff of intensive/extensive margins	ATE for $m(y) = \begin{cases} \log(y) & y > 0 \\ -x & y = 0 \end{cases}$	Pro: Explicit tradeoff of two margins Con: Need to choose x; Monotone only if support excludes $(0, e^{-x})$
Intensive margin effect	$E\left[\log\left(\frac{Y(1)}{Y(0)}\right) \mid Y(1) > 0, Y(0) > 0\right]$	Pro: ATE in logs for the intensive margin Con: Partially identified

Table 3: Table 2 from Chen and Roth (2023)

Comment 2 (Poisson Regression)

Poisson regression is a good example of a way to estimate $E(Y_i(1) - Y_i(0))/E(Y_i(0))$*. This approach estimates* $\log(E(Y|X)) = X\beta$*, rather than* $E(\log(Y)|X)$*. You get a simple semi-elasticity measure for the parameters, and* Y *can be zero. What are the typical concerns?*

1. *If* $Y|X$ *is truly distributed Poisson, conditional on* X*, then* $Var(Y|X) = E(Y|X)$*. This just comes from the Poisson distribution's statistical properties, but feels like a restrictive model assumption. But, it's not relevant for the parameter estimates of* β*. The estimates are still consistent, and the standard errors for these estimates can be adjusted for misspecification using robust standard errors (e.g. sandwich covariance estimators). These will give correct coverage, obviating any concerns about the Poisson regression. It is* not *necessary to use a Negative Binomial regression.*

2. *As we will discuss shortly, in many non-linear models, if you include parameters, such as fixed effects, which cannot be consisitently estimated, then this will make* all *the estimates in the model inconsistent. This is different from linear models. This concern is less of an issue in Poisson regression, as fixed effects can be concentrated out (see* `PPMLHDFE` *in Stata and* `glmhdfe` *in R)*

3. *Individuals are often not sure how to do instrumental variables in Poisson regression, but it is feasible! See Mullahy [1997], Windmeijer and Santos Silva [1997].*

The benefits of using the Poisson model (instead of log(1+Y)) according to Cohn et al. [2022]: "We replicate data sets from six papers published in top finance journals that together study two count or count-like outcomes... We...estimate log1plus and Poisson regressions based on that specification, and compare the coefficients of interest. These coefficients differ markedly in all six cases and have different signs in three of the six, suggesting that inference about even the direction of a relationship is sensitive to regression model choice in real-world applications...in all five cases involving regressions with control variables, switching from a log1plus to Poisson regression results in a larger change in the coefficient of interest than omitting the most important control variable, generally by a wide margin."

Inconsistency in binary choice models

Consider estimating a panel fixed effects model with binary choice:

$$\begin{aligned} Y_{it} &= \alpha_i + X_{it}\beta + \epsilon_{it} \\ Y_{it} &= F(\alpha_i + X_{it}\beta) \end{aligned}$$

where we are interested in the parameter β. If we have a short panel (e.g. few time periods), we cannot consistently estimate α_i. However, in the linear case, this does not affect estimation of β. More shockingly, however, is that for binary outcome case, the only model that consistently estimates β is a conditional Logit [Chamberlain, 1980, 2010].

More generally, if you have inconsistent fixed effects in your nonlinear models, this can cause serious issues (except in special cases like conditional). Often, the only way to get around these issues is by finding ways to "concentrate" or get around these nuisance parameters. Famous cases where this occurs include conditional logit, Poisson unit fixed effects, and partial likelihoods in the Cox proportional hazard model.

Multiple Choices

We'll now examine multiple discrete choice problems. Much of this discussion is very adjacent to industrial organization. However, many of these ideas are important for non-IO problems, such as multiple IVs and Roy models. Moreover, these tools are very promising in fields that have not yet used them.

Issues with choice problems that we'll discuss:

- Independence of Irrelevant Alternatives (IIA)
- Choice sets and consideration sets

Consider the following problem: we observe choices for individuals $Y_i = j$, $j \in \Omega = \{0, 1, \ldots, J\}$, where $J + 1 = |\Omega|$ is the total number of choices. Importantly, the order of the choices has no particular meaning. This could be red bus, blue bus and car as transportation choices, for example.

Given these sets of choices, we have different types of covariates we can observe. Some characteristics are choice specific (such as a price), while some are unit specific (such as a person's income). Often, we want to allow for the characteristics to vary by both dimensions. This includes allowing for a choice's characteristic to vary depending on the person (e.g. a unit specific coefficient on the choice's characteristic), or allowing the person's characteristic to have differential effects on the choice of different goods. In total, we have three types of characteristics:

1. X_i (individual characteristics, invariant to choices),
2. X_j (choice characteristics)
3. X_{ij} includes individual-by-choice characteristics

We can write X_i as X_{ij} by interacting with choice fixed effects, and X_j can have i speicfic coefficients.[13]

Now recall there are two (non-exclusive) ways to think the discrete choice problem. The first is a statistical view: namely, how do we model the choice probabilities? In the binary choice problem, there is only one parameter that needs be known, conditional on X_i: $\pi(X_i) = Pr(Y_i = 1|X_i)$ With more than two choices, the dimensionality becomes more complicated. We now have $\pi_j(\mathbf{X}), j = 2, 3$ for 3 choices.

How should we parameterize how other choices' characteristics affect each other? Most of the models we will discuss will make very specific restrictions on how choices affect one another. These are not innocuous choices, as we'll see, but they provide a huge amount of additional structure that can be used to identify the parameters of interest.

The naive approach

If we want to estimate simple treatment effects, we could focus on binary outcomes. For exmaple: we have a randomly assigned treatment T, and J choices. What is the effect of T on $Pr(Y_i = j)$ under random assignment?

$$\tau_j = Pr(Y_i = j|T_i = 1) - Pr(Y_i = j|T_i = 0) \tag{4}$$

The downisde of this approach is that there's no information about the substitution patterns of individuals in this form. Concretely, if τ_2 is positive, is that because the share of individuals choosing $Y_i = 1$ decreases, the share of individuals choosing $Y_i = 0$ decreases, or both? Namely, what is the *substitution* pattern across the choices?[14]

Nonetheless, it is still very helpful to estimate these measures, and it's useful when faced with a lot of choices to focus on the effect on one margin. We will need more structure to estimate relative choice substitution across outcomes, and ask questions like "what is the effect of T on choosing j conditional on choosing j or k?"

[13] Note that when $J = 1$, we collapse down to binary choice.

[14] To put a statistical note on this, there are effectively two endogenous variables ($1(Y_i = 1)$ and $1(Y_i = 2)$), and we only have one randomly assigned variable (T). Hence, there's no way to simultaneously identify the effect on both.

Conditional logit

A second way to view the problem is as an structural (economic) choice problem (pioneered by McFadden [McFadden, 1972]). Consider a set of utilities U_{ij} (unobserved) such that

$$Y_i = \arg\max_{j \in \Omega} U_{ij}. \tag{5}$$

In other words, person i chooses j if it's the choice that maximizes the utility amongst all $J+1$ choices. Note the similarity to the Y_i^* in the binary case!

If we make the assumptions:

1. $U_{ij} = X_{ij}'\beta + \varepsilon_{ij}$
2. ϵ_{ij} are independent across choices and individuals, and distributed Type-I extreme value

then we get the McFadden conditional logit model:

$$Pr(Y_i = j|X_i) = \frac{\exp(X_{ij}\beta)}{\sum_{k=0}^{J}\exp(X_{ik}\beta)}. \tag{6}$$

Comment 3
Note that if the characteristics X_{ij} only vary based on the individual (e.g. we can write $X_{ij}\beta$ as $X_i\beta_j$), then the effects across choices are relative *to each other. We can write our probability equation as*

$$Pr(Y_i = j|X_i) = \frac{\exp(\alpha_j + X_i\beta_j)}{1+\sum_{k=1}^{J}\exp(\alpha_k + X_i\beta_k)}. \tag{7}$$

This is the multinomial logit. *Once we allow for choice speicfic characteristics, then we need to write the probability following Equation (6).*

In many choice problems, a key parameter we're interested in is the price elasticity. The definition of the price elasticity is the percentage change in a market share of a good for a given percentage change in the price. Formally, the own-price elasticity is:

$$\epsilon_j = \frac{\partial Pr(Y_i = j|X_{ij})}{Pr(Y_i = j|X_{ij})}\frac{p_j}{\partial p_j} = \frac{\partial Pr(Y_i = j|X_{ij})}{\partial p_j}\frac{p_j}{Pr(Y_i = j|X_{ij})}. \tag{8}$$

We can *also* think about *cross*-price elasticities, e.g. how do market shares change when other goods shift their price:

$$\epsilon_{jk} = \frac{\partial Pr(Y_i = j|X_{ij})}{\partial p_k}\frac{p_k}{Pr(Y_i = j|X_{ij})}. \tag{9}$$

Note that with equation (6) as our probability model, we can estimate all these elasticities (assumign we have the data on prices, and we are willing to assume prices are exogeneous, a very strong assumption). But, this formulation creates issues.

A key issue with this formulation of the conditional logit model is that the cross-price elasticities are identical. Specifically, $\epsilon_{jk} = \epsilon_{lk}$, such that the effect of shifting price of a different good causes an identical proportionate shift in all choices' market share. You can see this by simply plugging in for $\frac{\partial Pr(Y_i=j|X_{ij})}{\partial p_k}$:

$$\begin{aligned}\epsilon_{jk} &= \underbrace{-\gamma Pr(Y_i = j|X_{ij})Pr(Y_i = k|X_{ij})}_{\frac{\partial Pr(Y_i=j|X_{ij})}{\partial p_k}} \times \frac{p_k}{Pr(Y_i = j|X_{ij})} \\ &= -\gamma Pr(Y_i = k|X_{ij})p_k,\end{aligned}$$

where γ is the coefficient on price in the conditional logit model. Note that this elasticity is not a function of j, and hence identical for all other products.[15]

The canonical example of this is the "car, red bus and blue bus" example. Imagine a choice set where there are three choices for transportation: a car, and two busses: one red, and one blue. Presumably a person is purely indifferent between red and blue busses. Hence, a shift in the red bus price would presumably cause a bigger substitution from the blue bus than from car users, but the conditional logit (in this form) will not account for this.

How can we deal with the IIA issue? This is a problem of poor substitution patterns, which is an economics problem. In other words, economics gives us an intuition about the market substitution patterns, and we don't think identical cross-elasticities makes sense. It's also a statistical problem – there is a very strong statistical functional form we have assummed, which was analytically convenient but has somewhat perverse properties. We will now consider a few (but not all) solutions to the problem proposed in the literature.

Data structure

Before jumping into estimation approaches to improve on the substitition patterns, it is worth touching on the structure of the observed data. Crudely, we can think of there being two types of data we observe. In some cases, we observe the individual's choices, combined with information about the individuals, and the choices themselves.[16] Often, these individuals may be in the same market, or in different markets. In other cases, we observe the market shares of the choices, the overall characteristics of the individuals in the market, and the characteristics of the choices.[17]

[15] It's useful to note that the *levels* of the market share do vary by good, but the elasticity scaling makes the cross-price elasticities identical.

[16] Sometimes we even observe their second and third choices, which can be very useful!

[17] Note that we could construct the market shares directly from the individual data, assuming that it's a random sample of the population.

Note that the market share implied by the model in Equation (6) is just $s_j \equiv Pr(Y_i = j|X_i)$. We can then think about transformations of this market share: $\log(s_j) = X_{ij}\beta - \log(\sum_{k=0}^{J} \exp(X_{ik}\beta))$.

Nested Logit and Correlated Multivariate Probit

One part of the IIA problem comes from the independence of ε across choices. Recall that the ε effectively rationalize the market shares beyond what we observe that is explained based on the covariates. Recall the blue and red bus case: getting two independent ε draws for the busses is not an intuitive view of bus demand. Instead, the blue and bus likely have highly correlated epsilon draws (if not identical), e.g. the unobserved latent demand for blue and red busses is correlated! The issue is exactly how to specify the correlation that preserves the ability to estimate the model.

With the nested Logit approach, you can specify sets (as the researcher), and allow correlation of the ε within these sets. The key is that the errors are uncorrelated across choice sets, which preserves the logit structure (see Goldberg [1995] for an example application), and the correlation *within* a nest is allowed to be correlated following a distinct similarity parameter. In essense, the similarlity parameter scales up and down the effect of the covariates within a nest: if the similarity is high, then the effect of the covariates is swamped by the random error, and the choices are highly correlated; if the similarity is low, the nest approaches the standard IIA setting. See Wen and Koppelman [2001] for a more recent discussion.

An alternative approach is to allow the covariance matrix of the error terms to be flexibly estimated by the data using a multivariate normal:

$$\epsilon_i = (\epsilon_{i0}, \epsilon_{i1}, \ldots, \epsilon_{iJ}) \sim \mathcal{N}(0, \Sigma) \tag{10}$$

where the researcher will then directly estimate Σ. Unfortunately, this problem gets hard with many choices (parameter space grows at rate $O(J^2)$). See McCulloch et al. [2000] and Geweke et al. [2003] for details and an application in the Bayesian setting, and Train (2009) for simulation discussions in the frequentist case.

Rather than directly target the distribution of the ε_{ij}, an alternative approach is to add more richness to the coefficients themselves. By adding more random variation in the loadings, it effectively creates a richer substitution pattern by adding more to the error term. Consider a slight extension of our previous model, with β_i varying by individual (in an unobserved way):[18]

[18] Note that this random variation in preferences is usually viewed as *exogeneous*.

$$
\begin{aligned}
U_{ij} &= X_{ij}\beta_i + \varepsilon_{ij} \\
U_{ij} &= X_{ij}\overline{\beta} + \nu_{ij}, \qquad \nu_{ij} = \varepsilon_{ij} + X_{ij}(\beta_i - \overline{\beta})
\end{aligned}
$$

There are a number of ways to estimate this approach, but notice the key point – subtitution patterns are more richly modeled (and allowed) due to ν_{ij} varying by X_{ij}.

Example 2 (Random coefficients estimation example)
Let $J = 3$, and X_j be a scalar (e.g. price). We assume that

$$U_{ij} = X_j \beta_i + \varepsilon_{ij} \qquad \beta_i = (\overline{\beta} + \sigma \nu_i), \nu_i \sim \mathcal{N}(0,1). \tag{11}$$

Separate the utility of choosing j into

$$U_{ij} = \mu_{ij}(\overline{\beta}) + X_j \sigma \nu_i + \varepsilon_{ij} \tag{12}$$
$$\mu_{ij} = X_j \overline{\beta}. \tag{13}$$

We can write the probability of choosing j as:

$$Pr(Y_i = j | X, \overline{\beta}, \sigma) = \int \frac{\exp(X_j \overline{\beta} + X_j \sigma \nu_i)}{\sum_{k=0}^{J} \exp(X_j \overline{\beta} + X_j \sigma \nu_i)} \phi(\nu_i) d\nu_i \tag{14}$$

where $\phi(\cdot)$ is the Normal standard normal pdf.
This setup is often referred to as a "mixed logit" model (in contrast with the more common Berry Levinsohn Pakes approach, which we'll discuss later) [McFadden and Train, 2000]. The typical approach for estimating these models involves using Maximum Simulated Likelihood, or Method of Simulated Moments. McFadden and Train [2000] show that a straightforward approach to estimating this is to simulate the model S times, and then use the simulated data to approximate the integral:

$$\hat{E}(Pr(Y_i = j | X, \overline{\beta}, \sigma)) = \frac{1}{S} \sum_{s=1}^{S} \frac{\exp(X_j \overline{\beta} + X_j \sigma \nu_{is})}{\sum_{k=0}^{J} \exp(X_j \overline{\beta} + X_j \sigma \nu_{is})}. \tag{15}$$

Then, this probability can be used to form a log-likelihood function, and the model can be estimated using standard optimization techniques for maximizing log-likelihoods.
Note that an important piece in this setting is micro-level choice data (which we use to form the likelihood), and the lack of any unobserved heterogeneity that creates endogeneity and bias in our estimates. Without an additional error term, there's no need for an instrument here. This is a version of assuming exogeneity conditional on observables. Often, we will only observe market-level *shares of goods. Then, we'll need many markets in order to have sufficient independent variation to estimate parameters. We will discuss this next.*

The workhorse set of demand estimation models is known as BLP (Berry Levinsohn Pakes), named after the authors in Berry et al. [1995]. This model combines random coefficient estimation with unobserved market-good-level demand heterogeneity that is potentially endogeneous and correlated with price. In other words, not only are individuals allowed to have random (independent) error, but there is a fixed unobserved error in demand for each good. This allows for a highly correlated set of demand choices within a market, and also creates unobserved demand heterogeneity that requires an instrument.

This model is often specified using the following utility function:

$$U_{ijm} = \delta_{jm} + \mu_{ijm} + \epsilon_{ijm}, \tag{16}$$

where $\delta_{jm} = X_j\beta + \xi_{jm}$ is the mean utility of choosing j in market m, μ_{ijm} is the random substition pattern specific to an individual (typically driven by the random coefficients on good characterstics as in Example 2), and ϵ_{ijm} is the individual specific logit error that is i.i.d. Often, this type of setting is used when only market-level data is available, and so the researcher observes the market shares of goods, but not the individual choices.[19]

Under the standard logit distributional assumptions for ϵ_{ijm},

$$Pr(Y_{im} = j|X) = s_{jm}(\boldsymbol{\delta}_m, \theta) = \int \frac{\exp(\delta_{jm} + \mu_{ijm})}{\sum_{k \in J_m} \exp(\delta_{km} + \mu_{ikm})} f(\boldsymbol{\mu}|\theta) d\boldsymbol{\mu}_{im}. \tag{17}$$

The key insight in Berry et al. [1995] is to note that the vector of δ_{jm} in market m, δ_m, can be inverted from the market shares, s_m, and θ, the parameters of the random mixing coefficients. Once we know δ_m, we can define $\xi_{jm} \equiv \delta_{jm} - X_j\beta$, and define a conditional moment condition $E(\xi_{jm}|Z_{jm}) = 0$. This moment condition can be used to estimate β using GMM. Conlon and Gortmaker [2020] provide a very nice discussion of the algorithmic approach on how to do this, and provide a Python package to solve this problem.[20]

Conclusion

Underlying structure of discrete choice is valuable in IV settings. Much of this discussion centered on IO style applications. But this discussion shows up when thinking about Roy style models.[21] When we discuss instruments and individuals' choice to take up a policy or not, if the policy is multi-dimensional, this types of models play a huge role. Recall our discussion of propensity scores for treatment effects. If individuals choose between multiple treatment options, this maps directly into a discrete choice setting like what we've discussed today. Thinking carefully about the counterfactual pattern across will give guidance in more complicated IV settings.

There is also value in arbitraging IO methods in other fields. Many fields have discrete choice applications but have not adopted the tools. The cutting edge of IO tools is quite complex, but this type of

[19] This is a common setting in many IO applications, where the researcher observes the market shares of goods, but not the individual choices. However, it's wonderful when you have more, and a host of papers using the micro data exist as well [Berry et al., 2004, Conlon and Gortmaker, 2023].

[20] Part of the reasoning for this is that the trick to invert the shares and recover δ_{jm} is a non-linear fixed point problem that needs to converge to a high degree of precision for successful estimation. Conlon and Gortmaker [2020] highlight the best approaches.

[21] See Hull [2018] for an example.

structure is very valuable when thinking about complicated choice patterns. Worthwhile to try to arbitrage these methods in fields that are less exposed to them (e.g. Koijen and Yogo [2019]).

References

Ghislain BD Aihounton and Arne Henningsen. Units of measurement and the inverse hyperbolic sine transformation. *The Econometrics Journal*, 24(2):334–351, 2021.

Marc F Bellemare and Casey J Wichman. Elasticities and the inverse hyperbolic sine transformation. *Oxford Bulletin of Economics and Statistics*, 82(1):50–61, 2020.

Steven Berry, James Levinsohn, and Ariel Pakes. Automobile prices in market equilibrium. *Econometrica*, 63(4):841–890, 1995.

Steven Berry, James Levinsohn, and Ariel Pakes. Differentiated products demand systems from a combination of micro and macro data: The new car market. *Journal of political Economy*, 112(1):68–105, 2004.

Gary Chamberlain. Analysis of covariance with qualitative data. *The review of economic studies*, 47(1):225–238, 1980.

Gary Chamberlain. Binary response models for panel data: Identification and information. *Econometrica*, 78(1):159–168, 2010.

Jiafeng Chen and Jonathan Roth. Logs with zeros? some problems and solutions. *The Quarterly Journal of Economics*, page qjad054, 2023.

Raj Chetty, Nathaniel Hendren, Patrick Kline, and Emmanuel Saez. Where is the land of opportunity? the geography of intergenerational mobility in the united states. *The Quarterly Journal of Economics*, 129(4):1553–1623, 2014.

Jonathan B Cohn, Zack Liu, and Malcolm I Wardlaw. Count (and count-like) data in finance. *Journal of Financial Economics*, 146(2): 529–551, 2022.

Christopher Conlon and Jeff Gortmaker. Best practices for differentiated products demand estimation with pyblp. *The RAND Journal of Economics*, 51(4):1108–1161, 2020.

Christopher Conlon and Jeff Gortmaker. Incorporating micro data into differentiated products demand estimation with pyblp. Technical report, NYU working paper, 2023.

John Geweke, Gautam Gowrisankaran, and Robert J Town. Bayesian inference for hospital quality in a selection model. *Econometrica*, 71 (4):1215–1238, 2003.

Pinelopi Koujianou Goldberg. Product differentiation and oligopoly in international markets: The case of the us automobile industry. *Econometrica: Journal of the Econometric Society*, pages 891–951, 1995.

Peter Hull. Estimating hospital quality with quasi-experimental data. *Available at SSRN 3118358*, 2018.

Ralph SJ Koijen and Motohiro Yogo. A demand system approach to asset pricing. *Journal of Political Economy*, 127(4):1475–1515, 2019.

Willard G Manning and John Mullahy. Estimating log models: to transform or not to transform? *Journal of Health Economics*, 20(4):461–494, 2001. ISSN 0167-6296. DOI: https://doi.org/10.1016/S0167-6296(01)00086-8. URL https://www.sciencedirect.com/science/article/pii/S0167629601000868.

Robert E McCulloch, Nicholas G Polson, and Peter E Rossi. A bayesian analysis of the multinomial probit model with fully identified parameters. *Journal of econometrics*, 99(1):173–193, 2000.

Daniel McFadden. Conditional logit analysis of qualitative choice behavior. 1972.

Daniel McFadden and Kenneth Train. Mixed mnl models for discrete response. *Journal of applied Econometrics*, 15(5):447–470, 2000.

John Mullahy. Instrumental-variable estimation of count data models: Applications to models of cigarette smoking behavior. *Review of Economics and Statistics*, 79(4):586–593, 1997.

John Ashworth Nelder and Robert WM Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 135(3):370–384, 1972.

Martin Ravallion. A concave log-like transformation allowing non-positive values. *Economics Letters*, 161:130–132, 2017.

Chieh-Hua Wen and Frank S Koppelman. The generalized nested logit model. *Transportation Research Part B: Methodological*, 35(7): 627–641, 2001.

Frank AG Windmeijer and Joao MC Santos Silva. Endogeneity in count data models: an application to demand for health care. *Journal of applied econometrics*, 12(3):281–294, 1997.

Lecture 1 - Potential Outcomes, Directed Acylic Graphs, and Structural Models

Paul Goldsmith-Pinkham

January 18, 2024

Not every economics research paper is estimating a causal quantity. But, the implication or takeaway of papers is (almost) always a causal one. Causality lies at the heart of every exercise. [1]

The goals in this lecture are:

- Enumerate tools used to discuss causal questions
- Emphasize a *multimodal* approach
- Set terminology/definitions for future discussions

Concretely, this involves covering three ways, notationally, of considering causal questions:

1. the potential outcomes (PO) framework,
2. the directed acyclic graph (DAG) framework,
3. structural models.

Over the course of describing these, we will also refresh our memories on the difference between the estimator, the estimand and the estimate, and learn the identification condition for the average treatment effect (ATE).

Notation

We will begin by outlining some notation for potential outcomes. When defining treatment effects, this notation is extremely convenient and clear, particularly when considering settings with significant unobserved heterogeneity. However, since so much of the extant literature in economics (and econometrics) is written using more standard structural equations (e.g. $Y = X\beta + \epsilon$), it is important to be able to translate between the two. For the sake of completeness, I also want to expose you to the directed acyclic graph (DAG) framework, [Pearl, 2009, Imbens, 2020] which is more commonly used in other fields such as epidimiology and computer science.[2] It is much less common in economics, but without getting into broader epistemic debates, it's extremely useful in some settings for clarifying the identifying assumptions (especially in settings relying on a "conditional on observables" assumption).

[1] "We do not have knowledge of a thing until we have grasped its why, that is to say, its cause." – Aristotle

[2] There was a period of time when the debate about DAGs was quite ornery (especially online). I think this has subsided.

Potential Outcomes

We will follow Imbens and Rubin [2015] in our notation. I will be slightly looser in my definitions for the purpose of space, but I encourage you to read Chapter 1 of Imbens and Rubin [2015] or Chapter 7 of Aronow and Miller [2019] for a more precise treatment.

Consider a sample of N units, indexed by i. Each unit has a treatment status D_i and an outcome Y_i.[3] Sometimes, I will refer to the collection of observations or treatments as $\mathbf{D}$ and $\mathbf{Y}$ to denote a vector of length N with each element corresponding to the treatment or outcome for a given unit. Both $\mathbf{D}$ and $\mathbf{Y}$ are *observed* in our data: we see who is treated ($\mathbf{D}$), and the subsequent outcome (the $\mathbf{Y}$ given the $\mathbf{D}$)

Example 1
Many medical examples naturally lend themselves to thinking about potential outcomes. For example, consider the outcome of whether you have a headache in three hours:

$$Y = \begin{cases} 1 & \textit{Have a headache in three hours} \\ 0 & \textit{Do not have a headache in three hours} \end{cases}$$

and the treatment of taking an aspirin:

$$D = \begin{cases} 1 & \textit{Take an aspirin} \\ 0 & \textit{Do not take an aspirin.} \end{cases}$$

We now consider the *potential* outcome for unit i. We can denote this as $Y_i(\mathbf{D})$, which is the outcome for unit i if the set of treatments for the N units is $\mathbf{D}$. Note that this is a complicated function! It depends on the treatment status of all units, not just the treatment status of unit i. This leads us to a first important assumption:

Assumption 1 (Stable Unit Treatment Value Assumption)
If $D_i = D_i'$, then $Y_i(\mathbf{D}) = Y_i(\mathbf{D}')$.

Put in words, it means that your potential outcome is only affected by your own treatment status, and not the treatment status of others.[4] This assumption lets us write our potential outcome as $Y_i(D_i)$, and focus just on how our own treatment affects our outcome. This is a strong assumption; we will discuss how one might consider relaxing it in a few lectures. And of course any macroeconomist will tell you that this is a terrible assumption. But, it is a useful starting point.

[3] For now, we will assume that there is just a binary treatment, but this can be generalized to multiple or continuous treatments. It will make life more complicated.

[4] Sometimes this is called a "no interference" condition. As we'll see later on, this could also be labeled a spillover in the economics literature.

Example 1 (continued)
We can now consider the potential *outcome in the state of the world where a person takes an aspirin or not: $Y_i(1)$ vs. $Y_i(0)$. Note that it is not fundamentally possible to observe both states of the world: even if a person were observed in different time periods, and in one case they took the aspirin and in another they did not, this would reflect fundamentally different observations. This type of repeated observation could be used to help identify the average potential outcomes, but would require additional assumptions.*
SUTVA is a very natural assumption in our medical example, since others' aspirin treatment decision should have no impact on our headache. However, this is likely not true with vaccines or other interventions.

It's worth remarking on a few things. First, this potential outcome is an function of the individuals' treatment status, and allowed to vary by individual. Second, this outcome itself is not necessarily observed. Indeed, what we observe is

$$Y_i = Y_i(1)D_i + Y_i(0)(1 - D_i). \qquad (1)$$

Hence, for the untreated units, we observe their $Y_i(0)$, and vice versa for the treated units. This model is often referred to as the Neyman-Rubin Causal model.[5]

The fact that we only observe either $Y_i(1)$ or $Y_i(0)$ is sometimes called the "fundamental problem of causal inference.[6]" Since we can only observe one outcome for a given unit, we cannot trace out the counterfactual outcomes for a single unit. This makes it quite challenging to know what the effect of changing D_i is on a single unit i.[7]

One way to view the fundamental problem of causal inference is as a missing data problem.[8] We will use many different techniques throughout this course to impute a counterfactual outcome such that we can know the causal effect of an intervention.

[5] These were not coauthors - Jerzy Neyman was a Polish statistician who initially proposed the potential outcomes framework to study completely randomized experiments [Splawa-Neyman et al., 1990]. This model was adopted and expanded by Donald Rubin in a number of influential papers. This model was coined the Rubin Causal model by Paul Holland, in an influential paper [Holland, 1986] about statistics and causality that we will revisit shortly.

[6] This term, again, comes from Holland [1986] (which you should read!).

[7] If we assume the treatment effects everyone exactly the same, then it is straightforward. While we might make a homogeneity assumption like this, we don't always believe it in practice.

[8] The treatment by Aronow and Miller [2019] covers this in very nice detail.

Comment 1

*It is worth thinking a bit about what causal effect you are interested in estimating. Often this is referred to as the **estimand**. This could be many things:*

- *A structural parameter (dInvestment/dTaxRate?)*
- *The effect of zoning restrictions on housing supply*
- *A policy evaluation of a renter's assistance program*
- *The existence of underreaction in stock prices to earnings news*

Comment 2

It is important to get these terms straight.

- *Estimand: the quantity to be estimated*
- *Estimate: the approximation of the estimand using a finite data sample*
- *Estimator: the method or formula for arriving at the estimate for an estimand*

For a particularly goofy way to remember this: `https://twitter.com/paulgp/status/1275135175966494721?s=20`

Identification of the Average Treatment Effect Estimand

We will conclude this lecture by describing sufficient conditions under which we can identify the **Average Treatment Effect** or ATE, a common target estimand for researchers.

Before we do that, we need to define the individual level causal estimand (that is, recall, inherently unknowable). Call this the **Individual Treatment Effect** or ITE. This is the difference between the potential outcomes for a given unit:

$$\tau_i \equiv Y_i(1) - Y_i(0). \tag{2}$$

This can be easily generalized to multiple treatments as well: we will discuss this in a few lectures.

Average Treatment Effect

We now consider the *average* treatment effect over the population. This is, quite simply, the average of the individual treatment effects over all individuals in the overall population.

Definition 1
We define the average treatment effect in our population as

$$\tau_{ATE} \equiv \mathbb{E}(\tau_i) = \mathbb{E}(Y_i(1) - Y_i(0)) = \mathbb{E}(Y_i(1)) - \mathbb{E}(Y_i(0)).$$

Why do we find the ATE interesting?[9] For one thing, it describes the effect of giving the treatment to everyone in the population. This is often of interest to policymakers, who want to know the effect of a policy on the entire population.

We now consider some additional average treatment estimands. The first is the Average Treatment Effect on the Treated (ATT):

Definition 2

$$\tau_{ATT} \equiv \mathbb{E}(\tau_i | D_i = 1) = \mathbb{E}(Y_i(1)|D_i = 1) - \mathbb{E}(Y_i(0)|D_i = 1).$$

This estimates the effect for individuals who received the treatment.[10] Note that one piece of the ATT is observed: $E(Y_i(1)|D_i = 1)$. This is just the observed outcome for the treated units.

We can also define the conditional average treament effect (CATE). Let X_i be a pre-determined set of covariates. Then, we can define the CATE as:

Definition 3

$$\tau_{CATE}(x) \equiv \mathbb{E}(Y_i(1)|X_i = x) - \mathbb{E}(Y_i(0)|X_i = x).$$

Example 1 (continued)
The ATE is what the effect would be on headaches if every person in the population took aspirin relative to not taking aspirin.
The ATT is what the impact of aspirin has been for those who took aspirin, relative to if they had not taken aspirin.
The CATE is what the impact of aspirin for a particular group, such as older men, would be relative to not taking aspirin.

It is useful to note the following relationship between the ATE and the CATE:

$$\tau_{ATE} = \int \tau_{CATE}(x) f(x) dx.$$

If X_i is discrete with values in $\mathcal{X}$, more simply this is

$$\tau_{ATE} = \sum_{x \in \mathcal{X}} \tau_{CATE}(x) Pr(X_i = x).$$

We now discuss under what conditions we can identify the ATE.

[9] This is sometimes called the *population average treatment effect* or PATE. This is then contrasted with the *sample average treatment effect* or SATE. The SATE is the ATE defined for the sample of N individuals we observe, while the PATE is the ATE for the population of individuals we can draw the sample from. We will discuss this in more detail later in the class, but for now I will just refer to the ATE to refer to the average treatment effect over the sample, and assume that the SATE and PATE are similar. Typically if samples are randomly draw, this is a reasonable assumption, and the difference is mainly in the inference. See Imbens [2004] for an example discussion.

[10] It will be a little while until we discuss cases when the ATT and ATE are different. A notable example is difference-in-differences. Many examples where we use *models* to estimate our counterfactual outcome will lead to cases where we can only identify the ATT and not the ATE.

Identification of the ATE

Comment 3
What is identification? Intuitively, for a given estimand to be identified, it means that in a world with no uncertainty about data, can we always identify the value of our estimand from the data we observe? To quote Lewbel [2019]: "Econometric identification really means just one thing: model parameters or features being uniquely determined from the observable population that generates the data."

Note that without further assumptions, the ATE is not identified from the observed data, $(\mathbf{Y}, \mathbf{D})$. Why? Consider the following estimator of the ATE:

$$\tau = E(Y_i | D_i = 1) - E(Y_i | D_i = 0) \tag{3}$$

which compares the treated units' average outcome to the untreated units' average outcome. Rewriting using our potential outcomes,

$$\tau = E(Y_i(1) | D_i = 1) - E(Y_i(0) | D_i = 0),$$

we see that our key challenge is that the two expectations condition on *different* values of D_i. Hence, if there is correlation between D_i and $(Y_i(1), Y_i(0))$, these two averages are not comparable.

Example 2
Imagine I am a researcher studying the effect of a wage training program (D_i) on wages (Y_i). I have a sample of workers, and I observe their wages and whether they participated in the training program. I want to know the effect of the training program on wages. If I use Equation (3) to compare the wages of those who take the program to those who do not, I may be comparing individuals who are very different.

For example, if the training program is voluntary, then it is likely that those who take the program are more motivated or knowledgeable about the labor force, and hence would have higher wages even if they did not take the program. In this case, the naive estimator would overstate the effect of the training program on wages. Let U_i be a binary variable capturing their motivation or knowledge of the labor force. If $E(D_i | U_i = 1) - E(D_i | U_i = 0) > 0$, and $E(\tau_i | U_i = 1) - E(\tau_i | U_i = 0) > 0$, then the naive estimator will overstate the effect of the training program on wages.

Comment 4
As an exercise, prove that the naive estimator is biased in Example 2.

We are now ready for our first identification result. We first define **strong ignorability**:

Definition 4
We say that D_i is strongly ignorable conditional on a vector $\mathbf{X}_i$ if

1. $Y_i(0), Y_i(1) \perp D_i | \mathbf{X}_i$
2. $\exists \varepsilon > 0$ *such that* $\varepsilon < Pr(D_i = 1|\mathbf{X}_i) < 1 - \varepsilon$.

The first part of Definition 4 is sometimes referred to unconfoundeness (or in economics, exogeneity): we assume that the choice of treatment is independent (conditional on $\mathbf{X}$) of the units' potential outcome. This means a unit can't select into the treament based on their potential benefits.[11]

The second condition asserts that there is some variation in treatment. This is sometimes called the common support or overlap condition . It is a bit stronger than we need, but it is a convenient way to ensure that we can compare the treated and untreated units.

Theorem 1 (Identification of the ATE)
If D_i is strongly ignorable conditional on $\mathbf{X}_i$, then

$$\mathbb{E}(\tau_i) = \sum_{x \in \text{Supp } X_i} \Big(\mathbb{E}(Y_i | D_i = 1, \mathbf{X}_i = x) - \mathbb{E}(Y_i | D_i = 0, \mathbf{X}_i = x) \Big) Pr(\mathbf{X}_i = x)$$

Proof 1
Note that by strong ignorability,

$$\mathbb{E}(Y_i(0)|\mathbf{X}_i) = \mathbb{E}(Y_i(0)|D_i = 0, \mathbf{X}_i) = \mathbb{E}(Y_i|D_i = 0, \mathbf{X}_i).$$

In essence, independence of D_i and $(Y_i(0), Y_i(1))$ lets us interchange counterfactuals and realized data in conditionals. The rest follows by the law of iterated expectations.

This result is quite powerful, and describes a non-parametric condition for when we can identify (and estimate) the ATE. A corrolary of this theorem is that we can also identify conditional average treatment effects as well (by assumption).

Identification through Directed Acylic Graphs

Above, we encoded random variables' relationships functionally, using potential outcomes. An alternative approach does this graphically. I will not cover this in significant detail, but want to give an

[11] Strong ignorability is a much more precise term than exogeneous, but tends to be used less in economics. When communicating with an economics audience, you might say that D_i is conditionally randomly assigned, or D_i is exogeneous – but this would omit the second condition (which is that the treatment is not too rare or too common).

example of how to think about identification using Directed Acylcic Graphs (DAGs).

We can encode the relationship between D and Y using an *arrow* in a graph. The direction emphasizes that D causes Y, and not vice versa.

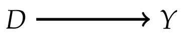


Figure 1: D has a causal effect on Y

We can also allow for the unobservable U, which drove identification concerns above in Example 2. In this case, U is termed a *confounder*. We can look at the paths by which D links to Y:

- The standard direct effect $D \to Y$
- The "back door" path $D \leftarrow U \to Y$

Note that the back-door is *not* causal. We know from above that the effect of D on Y is not identified under this setup, but this provides a graphical intuition as well – there is a path connecting D and Y but it does not flow in the right direction.

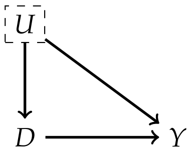


Figure 2: D's effect on Y is confounded by U

Now, we can replace U with an observable X. X is still a confounder, but since it is observable, we can condition on it and identify our effect (as in 1). As before, examine the paths by which D links to Y:

- The standard direct effect $D \to Y$
- The "back door" path $D \leftarrow X \to Y$.

In a DAG, conditioning on a variable along the path "blocks" the path, such that we would block the back door path.

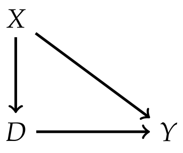
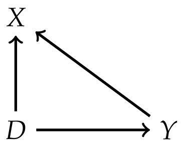


Figure 3: D's effect on Y is confounded by an observable X

Finally, let's consider a more complicated example. X is now a "collider", such that D and Y both affect X.

As before, examine the paths by which D links to Y:



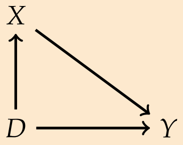
- The standard direct effect $D \to Y$
- The indirect path $D \to X \leftarrow Y$.

This is not called a backdoor path because X does not point into D.

The key difference in this setting is that since X does not cause Y, it is automatically blocked (all effects on X occur through our main effect). However, if you condition on X, you open the path!

Example 2 (continued)
Return to the example of a job training program. We want to study the impact of the program on wages, and we condition on whether a person has a car. If a person's wages affects their likelihood of having a car, we will have created a biased comparison: we will first consider the effect of the job training program among those who have a car (which may be small), and then among those who do not (which may also be small). If much of the effect of the program affects individuals' ability to buy a car, then we will underestimate the effect of the program.

Comment 5
A last example is what's called a mediator. *This is another variable that is affected by the treatment, and affects the outcome. In this case, we can think of the treatment as having two effects: a direct effect, and an indirect effect through the mediator.*



It is possible to control for a mediator in order to estimate only the direct effect – this is sometimes referred to as mediation analysis. However this is very *sensitive to functional form, and not recommended.*

I will not give an exhaustive approach on how to deal with DAGs for identification, but you can hopefully see that there is a great deal of intuitive value in writing down the DAG in some problems. This is particularly true when dealing with *colliders*.

Structural equations and causal effects

It is important to not lose sight of the fact that these should be estimates that inform our *economic* model. Since much of our background is traditionally in structural equations (that often map to economic models) it can often be more familiar to write out outcome equation as:

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

It is quite helpful to see how this maps back to the potential outcome framework:

$$\begin{aligned}
Y_i &= Y_i(0)(1 - D_i) + Y_i(1)D_i \\
&= Y_i(0) + \tau_i D_i \\
&= Y_i(0) + \tau D_i + (\tau_i - \tau)D_i \\
&= \underbrace{E(Y_i(0)|D_i = 0)}_{\alpha} + \underbrace{\tau}_{\beta} D_i + \underbrace{(\tau_i - \tau)D_i + (Y_i(0) - E(Y_i(0)|D_i = 0))}_{\varepsilon_i}
\end{aligned}$$

Consider now what $E(Y_i|D_i)$ will recover:

$$\begin{aligned}
E(Y_i|D_i = 1) &= \alpha + \tau + E(\varepsilon_i|D_i = 1) \\
E(\varepsilon_i|D_i = 1) &= (E(\tau_i|D_i = 1) - \tau) + E(Y_i(0)|D_i = 1) - E(Y_i(0)|D_i = 0) \\
E(Y_i|D_i = 0) &= \alpha + E(\varepsilon_i|D_i = 1) \\
E(\varepsilon_i|D_i = 0) &= 0.
\end{aligned}$$

So, we can see that we will recover the average treatment effect as β if D_i is randomly assigned (or strongly ignorable). This is a special case where the coefficient β in the linear regression case will give the average treatment effect, the constant will give the average for the untreated, and the error term will capture the rest. We will suffer from omitted variable bias if $E(Y_i(0)|D_i = 1) - E(Y_i(0)|D_i = 0) \neq 0$ – e.g. if there is selection into treatment based on your control potential outcome. Notice that if $E(\tau_i|D_i = 1) \neq \tau$, then we will also not estimate the ATE, but we will estimate the ATT.

More generally, however, it's just useful to see that there are one-to-one mappings between the potential outcome framework and structural regressions. In many ways, the potential outcome framework is helpful because it emphasizes the relevant counterfactual state more than many linear models.

Phil Haile has some lovely slides discussing the importance of structure in economics. One of the key issues he pushes back on is the idea where many applied researchers estimating treatment effects say they are being "model-free." In other words, rather than writing down a structural model and attempting to estimate something complicated with a functional form, they view their treatment effects as model-agnostic. This is sometimes referred to as the "reduced form".

What is the reduced form from a structural estimation perspective? Following Haile, a reduced form relationship is one where the endogeneous variable is a function of *exogeneous* variables and unobserved structural error terms. Exogeneous here means variables that satisfy the necessary independence assumptions with the structural error terms.

Example 3
Consider a supply and demand system:

$$\begin{aligned} Q_d &= D(P, X, U_d) \\ Q_s &= S(P, Z, U_s). \end{aligned}$$

These are simultaneous equations where the observed price we see in the market is the price where $Q_d = Q_s$. *Often, supply* (Q_s) *will be written in terms of price (which is a function of marginal cost):*

$$\begin{aligned} Q &= D(P, X, U_d) \\ P &= S(Q, Z, U_s). \end{aligned}$$

Since P *and* Q *are endogeneous, these are* structural *equations.*

The reduced form version of these equations would have the form

$$\begin{aligned} Q &= d(X, Z, U_d, U_s) \\ P &= s(X, Z, U_d, U_s). \end{aligned}$$

In economics, we may consider estimating the effect of price on quantity (e.g. a labor demand elasticity), which is a parameter in the structural demand equation. When we use instrumental variables and two-stage least squares (to be discussed further in a later class), the first stage *will be the reduce form, and the second stage is a structural model.*

Overall, it's important to remember that many of our estimation approaches imply a particular structural model. We may be approximating something more complicated, but we're typically making some kind of modeling decision.

Discussion Questions 1

1. *Consider the potential outcome framework in the context of individuals. We are thinking about annual earnings Y_i for an individual i. Often, we study the earnings gap between men and women. Is it reasonable to consider the potential outcome $Y_i(1)$ vs. $Y_i(0)$ for $D_i = 1$ when i is a woman vs. when i is a man?*

2. *Consider the linear model from above:*

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

When would we expected homoskedasticity to hold?

References

Peter M Aronow and Benjamin T Miller. *Foundations of agnostic statistics*. Cambridge University Press, 2019.

Paul W Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986.

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

Guido W Imbens. Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. *Journal of Economic Literature*, 58(4):1129–1179, 2020.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Arthur Lewbel. The identification zoo: Meanings of identification in econometrics. *Journal of Economic Literature*, 57(4):835–903, 2019.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Jerzy Splawa-Neyman, D. M. Dabrowska, and T. P. Speed. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5(4):465–472, 1990. ISSN 08834237. URL http://www.jstor.org/stable/2245382.