STATS 361: Causal Inference

Stefan Wager
Stanford University

Spring 2020

Contents

Lecture 1
Randomized Controlled Trials

Randomized controlled trials (RCTs) form the foundation of statistical causal inference. When available, evidence drawn from RCTs is often considered gold statistical evidence; and even when RCTs cannot be run for ethical or practical reasons, the quality of observational studies is often assessed in terms of how well the observational study approximates an RCT.

Today's lecture is about estimation of average treatment effects in RCTs in terms of the potential outcomes model, and discusses the role of regression adjustments for causal effect estimation. The average treatment effect is identified entirely via randomization (or, by design of the experiment). Regression adjustments may be used to decrease variance, but regression modeling plays no role in defining the average treatment effect.

The average treatment effect We define the causal effect of a treatment via potential outcomes. For a binary treatment $w \in \{0, 1\}$, we define potential outcomes $Y_i(1)$ and $Y_i(0)$ corresponding to the outcome the i-th subject would have experienced had they respectively received the treatment or not. The causal effect of the treatment on the i-th unit is then[1]

$$\Delta_i = Y_i(1) - Y_i(0). \tag{1.1}$$

The fundamental problem in causal inference is that only one treatment can be assigned to a given individual, and so only one of $Y_i(0)$ and $Y_i(1)$ can ever be observed. Thus, Δ_i can never be observed.

[1]One major assumption that's baked into this notation is that binary counterfactuals exist, i.e., that it makes sense to talk about the effect of choosing to intervene or not on a single unit, without considering the treatments assigned to other units. This may be a reasonable assumption in medicine (i.e., that the treatment prescribed to patient A doesn't affect patient B), but are less appropriate in social or economic settings where network effects may arise. We will discuss causal inference under interference later in the course.

Now, although Δ_i itself is fundamentally unknowable, we can (perhaps remarkably) use randomized experiments to learn certain properties of the Δ_i. In particular, large randomized experiments let us recover the average treatment effect (ATE)

$$\tau = \mathbb{E}\left[Y_i(1) - Y_i(0)\right]. \tag{1.2}$$

To do so, assume that we observe n independent and identically distributed samples $(Y_i,\, W_i)$ satisfying the following two properties:

$$\begin{aligned} &Y_i = Y_i(W_i) && \text{(SUTVA)} \\ &W_i \perp\!\!\!\perp \{Y_i(0),\, Y_i(1)\} && \text{(random treatment assignment)} \end{aligned}$$

Then, the difference-in-means estimator

$$\hat{\tau}_{DM} = \frac{1}{n_1} \sum_{W_i=1} Y_i - \frac{1}{n_0} \sum_{W_i=1} Y_i, \qquad n_w = |\{i : W_i = w\}| \tag{1.3}$$

is unbiased and consistent for the average treatment effect

Difference-in-means estimation The statistical properties of $\hat{\tau}_{DM}$ can readily be established. Noting that, for $w \in \{0,\, 1\}$

$$\begin{aligned} \mathbb{E}\left[\frac{1}{n_w} \sum_{W_i=w} Y_i\right] &= \mathbb{E}\left[Y_i \,\middle|\, W_i = w\right] && \text{(IID)} \\ &= \mathbb{E}\left[Y_i(w) \,\middle|\, W_i = w\right] && \text{(SUTVA)} \\ &= \mathbb{E}\left[Y_i(w)\right], && \text{(random assignment)} \end{aligned}$$

we find that the difference-in-means estimator is unbiased[2]

$$\mathbb{E}\left[\hat{\tau}_{DM}\right] = \mathbb{E}\left[Y_i(1)\right] - \mathbb{E}\left[Y_i(0)\right] = \tau.$$

Moreover, we can write the variance as

$$\operatorname{Var}\left[\hat{\tau}_{DM} \,\middle|\, n_0,\, n_1\right] = \frac{1}{n_0} \operatorname{Var}\left[Y_i(0)\right] + \frac{1}{n_1} \operatorname{Var}\left[Y_i(1)\right].$$

A standard central limit theorem can be used to verify that

$$\begin{aligned} &\sqrt{n}\left(\hat{\tau}_{DM} - \tau\right) \Rightarrow \mathcal{N}\left(0,\, V_{DM}\right), \\ &V_{DM} = \operatorname{Var}\left[Y_i(0)\right] \,/\, \mathbb{P}\left[W_i = 0\right] + \operatorname{Var}\left[Y_i(1)\right] \,/\, \mathbb{P}\left[W_i = 1\right]. \end{aligned} \tag{1.4}$$

[2]For a precise statement, one would need to worry about the case where n_0 or n_1 is 0.

Finally, note that we can estimate V_{DM} via routine plug-in estimators to build valid Gaussian confidence intervals for τ:

$$\lim_{n\to\infty} \mathbb{P}\left[\tau \in \left(\hat{\tau}_{DM} \pm \Phi^{-1}(1-\alpha/2)\sqrt{\widehat{V}_{DM}/n}\right)\right] = 1-\alpha, \tag{1.5}$$

where Φ denotes the standard Gaussian cumulative distribution function and

$$\widehat{V}_{DM} = \frac{1}{n_1-1}\sum_{W_i=1}\left(Y_i - \frac{1}{n_1}\sum_{W_i=1} Y_i\right)^2 + \frac{1}{n_0-1}\sum_{W_i=0}\left(Y_i - \frac{1}{n_0}\sum_{W_i=0} Y_i\right)^2.$$

From a certain perspective, the above is all that is needed to estimate average treatment effects in randomized trials. The difference in means estimator $\hat{\tau}_{DM}$ is consistent and allows for valid asymptotic inference; moreover, the estimator is very simple to implement, and hard to "cheat" with (there is little room for an unscrupulous analyst to try different estimation strategies and report the one that gives the answer closest to the one they want). On the other hand, it is far from clear that $\hat{\tau}_{DM}$ is the "optimal" way to use the data, in the sense that it provides the most accurate value of τ for a given sample size. Below, we try to see if/when we can do better.

Example: The linear model To better understand the behavior of $\hat{\tau}_{DM}$, it is helpful to look at special cases. First, we consider the linear model: We assume that $(X_i,\, Y_i,\, W_i)$ is generated as

$$Y_i(w) = c_{(w)} + X_i\beta_{(w)} + \varepsilon_i(w),\;\; \mathbb{E}\left[\varepsilon_i(w)\,\middle|\, X_i\right] = 0,\;\; \text{Var}\left[\varepsilon_i(w)\,\middle|\, X_i\right] = \sigma^2. \tag{1.6}$$

Here, $\hat{\tau}_{DM}$ does not use the X_i; however, we can characterize its behavior in terms of the distribution of the X_i. Throughout our analysis, we assume for simplicity that we are in a balanced randomized trial, with

$$\mathbb{P}\left[W_i = 0\right] = \mathbb{P}\left[W_i = 1\right] = \frac{1}{2}.$$

Moreover, we assume (without loss of generality) that

$$\mathbb{E}\left[X\right] = 0, \quad \text{and define} \quad A = \text{Var}\left[X\right].$$

The assumption that $\mathbb{E}\left[X\right] = 0$ is without loss of generality because all estimators we will consider today are translation invariant (but of course the analyst cannot be allowed to make use of knowledge that $\mathbb{E}\left[X\right] = 0$).

Given this setup, we can write the asymptotic variance of $\hat{\tau}_{DM}$ as

$$\begin{aligned} V_{DM} &= \text{Var}\left[Y_i(0)\right] \big/ \mathbb{P}\left[W_i = 0\right] + \text{Var}\left[Y_i(1)\right] \big/ \mathbb{P}\left[W_i = 1\right] \\ &= 2\left(\text{Var}\left[X_i\beta_{(0)}\right] + \sigma^2\right) + 2\left(\text{Var}\left[X_i\beta_{(1)}\right] + \sigma^2\right) \\ &= 4\sigma^2 + 2\left\|\beta_{(0)}\right\|_A^2 + 2\left\|\beta_{(1)}\right\|_A^2 \\ &= 4\sigma^2 + \left\|\beta_{(0)} + \beta_{(1)}\right\|_A^2 + \left\|\beta_{(0)} - \beta_{(1)}\right\|_A^2, \end{aligned} \tag{1.7}$$

where we used the notation

$$\|v\|_A^2 = v'Av.$$

Is this the best possible estimator for τ?

Regression adjustments with a linear model If we assume the linear model (1.6), it is natural to want to use it for better estimation. Note that, given this model, we can write that ATE as

$$\tau = \mathbb{E}\left[Y(1) - Y(0)\right] = c_{(1)} - c_{(0)} + \mathbb{E}\left[X\right]\left(\beta_{(1)} - \beta_{(0)}\right). \tag{1.8}$$

This suggests an ordinary least-squares estimator

$$\hat{\tau}_{OLS} = \hat{c}_{(1)} - \hat{c}_{(0)} + \overline{X}\left(\hat{\beta}_{(1)} - \hat{\beta}_{(0)}\right), \quad \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \tag{1.9}$$

where the $(\hat{c}_{(w)},\, \hat{\beta}_{(w)})$ are obtained by running OLS on those observations with $W_i = w$ (i.e., we run separate regressions on treated and control units). Standard results about OLS imply that (recall that, wlog, we work with $\mathbb{E}\left[X\right] = 0$)

$$\sqrt{n_w}\left(\begin{pmatrix}\hat{c}_{(w)}\\ \hat{\beta}_{(w)}\end{pmatrix} - \begin{pmatrix}c_{(w)}\\ \beta_{(w)}\end{pmatrix}\right) \Rightarrow \mathcal{N}\left(0,\, \sigma^2\begin{pmatrix}1 & 0\\ 0 & A^{-1}\end{pmatrix}\right). \tag{1.10}$$

In particular, we find that $\hat{c}_{(0)}$, $\hat{c}_{(1)}$, $\hat{\beta}_{(0)}$, $\hat{\beta}_{(1)}$ and $\overline{X}$ are all asymptotically independent. Then, we can write

$$\begin{aligned} \hat{\tau}_{OLS} - \tau = &\underbrace{\hat{c}_{(1)} - c_{(1)}}_{\approx\mathcal{N}(0,\,\sigma^2/n_1)} - \underbrace{\hat{c}_{(0)} - c_{(0)}}_{\approx\mathcal{N}(0,\,\sigma^2/n_0)} + \underbrace{\overline{X}\left(\beta_{(1)} - \beta_{(0)}\right)}_{\approx\mathcal{N}\left(0,\,\left\|\beta_{(1)}-\beta_{(0)}\right\|_A^2/n\right)} \\ &+ \underbrace{\overline{X}\left(\hat{\beta}_{(1)} - \hat{\beta}_{(0)} - \beta_{(1)} + \hat{\beta}_{(0)}\right)}_{\mathcal{O}_P(1/n)}, \end{aligned}$$

which leads us to the central limit theorem

$$\sqrt{n}\left(\hat{\tau}_{OLS} - \tau\right) \Rightarrow \mathcal{N}\left(0,\, V_{OLS}\right), \quad V_{OLS} = 4\sigma^2 + \left\|\beta_{(0)} - \beta_{(1)}\right\|_A^2. \tag{1.11}$$

In particular, note that $V_{DM} = V_{OLS} + \left\|\beta_{(0)} + \beta_{(1)}\right\|_A^2$, and so OLS in fact helps reduce asymptotic error in the linear model.

Regression adjustments without linearity The above result is perhaps not so surprising: If we assume a linear model, than using an estimator that leverages linearity ought to help. However, it is possible to prove a much stronger result for OLS in randomized trials: OLS is never worse that the difference-in-means methos in terms of its asymptotic variance, and usually improves on it (even in misspecified models).

Replace our linearity assumption with the following generic assumption:

$$Y_i(w) = \mu_{(w)}(X_i) + \varepsilon_i(w), \quad \mathbb{E}\left[\varepsilon_i(w) \,\middle|\, X_i\right] = 0, \quad \text{Var}\left[\varepsilon_i(w) \,\middle|\, X_i\right] = \sigma^2, \quad (1.12)$$

for some arbitrary function $\mu_{(w)}(x)$. As before, we can check that (recall that we assume that $\mathbb{P}[W_i = 1] = 0.5$)

$$\sqrt{n}\,(\hat{\tau}_{DM} - \tau) \Rightarrow \mathcal{N}(0,\, V_{DM}) = 4\sigma^2 + 2\,\text{Var}\left[\mu_{(0)}(X_i)\right] + 2\,\text{Var}\left[\mu_{(1)}(X_i)\right],$$

and so $\hat{\tau}_{DM}$ provides a simple way of getting consistent estimates of τ.

In order to analyze OLS, we need to use the Huber-White analysis of linear regression. Without any assumption on $\mu_{(w)}(x)$, the OLS estimates $(\hat{c}_{(w)},\, \hat{\beta}_{(w)})$ converge to a limit characterized as

$$\left(c^*_{(w)},\, \beta^*_{(w)}\right) = \text{argmin}_{c,\,\beta}\left\{\mathbb{E}\left[(Y_i(w) - X_i\beta - c)^2\right]\right\}. \quad (1.13)$$

If the linear model is misspecified, $(c^*_{(w)},\, \beta^*_{(w)})$ can be understood as those parameters that minimize the expected mean-squared error of any linear model. Given this notation, it is well known[3] that (recall that we still assume wlog that $\mathbb{E}[X] = 0$)

$$\begin{aligned} &\sqrt{n_w}\left(\begin{pmatrix}\hat{c}_{(w)}\\ \hat{\beta}_{(w)}\end{pmatrix} - \begin{pmatrix}c^*_{(w)}\\ \beta^*_{(w)}\end{pmatrix}\right) \Rightarrow \mathcal{N}\left(0,\, \begin{pmatrix}MSE^*_{(w)} & 0\\ 0 & \cdots\end{pmatrix}\right) \\ &c^*_{(w)} = \mathbb{E}[Y_i(w)], \qquad MSE^*_{(w)} = \mathbb{E}\left[\left(Y_i(w) - X_i\beta^*_{(w)} - \hat{c}^*_{(w)}\right)^2\right] \end{aligned} \quad (1.14)$$

Then, following the line of argumentation in the previous section, we can derive a central limit theorem

$$\sqrt{n}\,(\hat{\tau}_{OLS} - \tau) \Rightarrow \mathcal{N}(0,\, V_{OLS}), \quad (1.15)$$

[3]For a recent review of asymptotics for OLS under misspecification, see Buja et al. [2019]; in particular (1.14) is stated as Proposition 7.1 of this paper.

with asymptotic variance[4]

$$
\begin{aligned}
V_{OLS} &= 2MSE^*_{(0)} + 2MSE^*_{(1)} + \left\|\beta^*_{(1)} - \beta^*_{(0)}\right\|^2_A \\
&= 4\sigma^2 + 2\,\mathrm{Var}\left[\mu_{(0)}(X) - X\beta^*_{(0)}\right] \\
&\quad + 2\,\mathrm{Var}\left[\mu_{(1)}(X) - X\beta^*_{(1)}\right] + \left\|\beta^*_{(1)} - \beta^*_{(0)}\right\|^2_A \\
&= 4\sigma^2 + 2\left(\mathrm{Var}\left[\mu_{(0)}(X)\right] - \mathrm{Var}\left[X\beta^*_{(0)}\right]\right) \\
&\quad + 2\left(\mathrm{Var}\left[\mu_{(1)}(X)\right] - \mathrm{Var}\left[X\beta^*_{(1)}\right]\right) + \left\|\beta^*_{(1)} - \beta^*_{(0)}\right\|^2_A \\
&= 4\sigma^2 + 2\left(\mathrm{Var}\left[\mu_{(0)}(X)\right] + \mathrm{Var}\left[\mu_{(1)}(X)\right]\right) \\
&\quad + \left\|\beta^*_{(1)} - \beta^*_{(0)}\right\|^2_A - 2\left\|\beta^*_{(0)}\right\|^2_A - 2\left\|\beta^*_{(1)}\right\|^2_A \\
&= 4\sigma^2 + 2\left(\mathrm{Var}\left[\mu_{(0)}(X)\right] + \mathrm{Var}\left[\mu_{(1)}(X)\right]\right) - \left\|\beta^*_{(0)} + \beta^*_{(1)}\right\|^2_A \\
&= V_{DM} - \left\|\beta^*_{(0)} + \beta^*_{(1)}\right\|^2_A .
\end{aligned}
$$

In other words, whether or not the true effect function $\mu_w(x)$ is linear, OLS always reduces the asymptotic variance of DM. Moreover, the amount of variance reduction scales by the amount by which OLS in fact chooses to fit the training data. A worst case for OLS is when $\beta^*_{(0)} = \beta^*_{(1)} = 0$, i.e., when OLS asymptotically just does nothing, and $\hat{\tau}_{OLS}$ reduces to $\hat{\tau}_{DM}$.

Recap The individual treatment effect $\Delta_i = Y_i(1) - Y_i(0)$ is central object of interest in causal inference. These effects Δ_i themselves are fundamentally unknowable; however, a large randomized controlled trial lets us consistently recover the average treatment effect $\tau = \mathbb{E}\left[\Delta_i\right]$. Moreover, even without assuming linearity, we found that OLS regression adjustments generally improve on the performance of the simple difference in means estimator.

We emphasize that, throughout our analysis, we defined the target estimand $\tau = \mathbb{E}\left[\Delta_i\right]$ *before* making any modeling assumptions. Linear modeling was only used as a tool to estimate τ, but did not inform the scientific question we tried to answer. In particular, we did *not* try to estimate τ by direct regression modeling $Y_i \sim X_i\beta + W_i\tau + \varepsilon_i$, while claiming that the coefficient on τ is a causal effect. This approach has the vice of tying our scientific question to our regression modeling strategy: τ appears to just have become a coefficient in our linear model, not a fact of nature that's conceptually prior to modeling decisions.

[4]For the third equality, we use the fact that $X\beta^*_{(w)}$ is the projection of $\mu_{(w)}(X)$ on to the linear span of the features X, and so $\mathrm{Cov}[\mu_{(w)}(X),\, X\beta^*_{(w)}] = \mathrm{Var}[X\beta^*_{(w)}]$.

Finally, note that our OLS estimator can effectively be viewed as

$$\hat{\tau}_{OLS} = \frac{1}{n}\sum_{i=1}^{n}\left(\underbrace{\left(\hat{c}_{(1)} + X_i\hat{\beta}_{(1)}\right)}_{\hat{\mu}_{(1)}(X_i)} - \underbrace{\left(\hat{c}_{(0)} + X_i\hat{\beta}_{(0)}\right)}_{\hat{\mu}_{(0)}(X_i)}\right), \tag{1.16}$$

where $\hat{\mu}_{(w)}(x)$ denotes OLS predictions at x. Could we use other methods to estimate $\hat{\mu}_{(w)}(x)$ rather than OLS (e.g., deep nets, forests)? How would this affect asymptotic variance? More on this in the homework.

Bibliographic notes The potential outcomes model for causal inference was first advocated by Neyman [1923] and Rubin [1974]; see Imbens and Rubin [2015] for a modern textbook treatment. Lin [2013] presents a thorough discussion of the role of linear regression adjustments in improving the precision of average treatment effect estimators. Wager, Du, Taylor, and Tibshirani [2016] have a discussion of non-parametric or high-dimensional regression adjustments in RCTs that expands on the results covered here.

One distinction question that has received considerable attention in the literature is whether or not one is willing to make any stochastic assumptions on the potential outcomes. In this lecture, we worked under a population model, i.e., we assumed the existence of a distribution P such that the potential outcomes are drawn as $\{Y_i(0),\, Y_i(1)\} \overset{\text{iid}}{\sim} P$, and we sought to estimate $\tau = \mathbb{E}_P\left[Y_i(1) - Y_i(0)\right]$. In contrast, others adopt a strict randomization inference framework where the potential outcomes $\{Y_i(0),\, Y_i(1)\}_{i=1}^n$ are taken as fixed, and only the treatment assignment W_i is taken to be a random variable; they then consider estimation of the sample average treatment effect $\tau_{SATE} = n^{-1}\sum_{i=1}^n (Y_i(1) - Y_i(0))$.

The advantage of the randomization inference framework is that it does not require the statistician to imagine the sample as a representative draw from a population; in contrast, the advantage of population modeling is that it often allows for simpler and often more transparent statistical arguments. The study of high-dimensional regression adjustments under randomization inference is an ongoing effort, with recent contributions from Bloniarz, Liu, Zhang, Sekhon, and Yu [2016] and Lei and Ding [2018].

Lecture 2
Unconfoundedness and the Propensity Score

One of the simplest extensions of the randomized trial is treatment effect estimation under unconfoundedness. Qualitatively, unconfoundedness is relevant when we want to estimate the effect of a treatment that is not randomized, but is as good as random once we control for a set of covariates X_i.

The goal of this lecture is to discuss identification and estimation of average treatment effects under such an unconfoundedness assumption. As before, our approach will be non-parametric: We won't assume well specification of any parametric models, and identification of the average treatment effect will be driven entirely by the design (i.e., conditional independence statements relating potential outcomes and the treatment).

Beyond a single randomized controlled trial We define the causal effect of a treatment via potential outcomes. For a binary treatment $w \in \{0, 1\}$, we define potential outcomes $Y_i(1)$ and $Y_i(0)$ corresponding to the outcome the i-th subject would have experienced had they respectively received the treatment or not. We assume SUTVA, $Y_i = Y_i(W_i)$, and want to estimate the average treatment effect

$$\text{ATE} = \mathbb{E}\left[Y_i(1) - Y_i(0)\right].$$

In the first lecture, we assumed random treatment assignment, $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp W_i$, and studied several $\sqrt{n}$-consistent estimators for the ATE.

The simplest way to move beyond one RCT is to consider two RCTs. As a concrete example, supposed that we are interested in giving teenagers cash incentives to discourage them from smoking. A random subset of $\sim 5\%$ of teenagers in Palo Alto, CA, and a random subset of $\sim 20\%$ of teenagers in Geneva, Switzerland are eligible for the study.

Palo Alto	Non-S.	Smoker
Treat.	152	5
Control	2362	122

Geneva	Non-S.	Smoker
Treat.	581	350
Control	2278	1979

Within each city, we have a randomized controlled study, and in fact readily see that the treatment helps. However, looking at aggregate data is misleading, and it looks like the treatment hurts; this is an example of what is sometimes called Simpson's paradox:

Palo Alto + Geneva	Non-Smoker	Smoker
Treatment	733	401
Control	4640	2101

Once we aggregate the data, this is no longer an RCT because Genevans are both more likely to get treated, and more likely to smoke whether or not they get treated. In order to get a consistent estimate of the ATE, we need to estimate treatment effects in each city separately:

$$\begin{aligned}
\hat{\tau}_{\text{PA}} &= \frac{5}{152+5} - \frac{122}{2362+122} \approx -1.7\%, \\
\hat{\tau}_{\text{GVA}} &= \frac{350}{350+581} - \frac{1979}{2278+1979} \approx -8.9\% \\
\hat{\tau} &= \frac{2641}{2641+5188}\hat{\tau}_{\text{PA}} + \frac{5188}{2641+5188}\hat{\tau}_{\text{GVA}} \approx -6.5\%.
\end{aligned}$$

What are the statistical properties of this estimator? How does this idea generalize to continuous x?

Aggregating difference-in-means estimators Suppose that we have covariates X_i that take values in a discrete space $X_i \in \mathcal{X}$, with $|\mathcal{X}| = p < \infty$. Suppose moreover that the treatment assignment is random conditionally on X_i, (i.e., we have an RCT in each group defined by a level of x):

$$\{Y_i(0),\, Y_i(1)\} \perp\!\!\!\perp W_i \,\big|\, X_i = x, \quad \text{for all} \quad x \in \mathcal{X}. \tag{2.1}$$

Define the group-wise average treatment effect as

$$\tau(x) = \mathbb{E}\left[Y_i(1) - Y_i(0) \,\middle|\, X_i = x\right]. \tag{2.2}$$

Then, as above, we can estimate the ATE τ by aggregating group-wise treatment effect estimations,

$$\hat{\tau}_{AGG} = \sum_{x \in \mathcal{X}} \frac{n_x}{n} \hat{\tau}(x), \quad \hat{\tau}(x) = \frac{1}{n_{x1}} \sum_{\{X_i = x,\, W_i = 1\}} Y_i - \frac{1}{n_{x0}} \sum_{\{X_i = x,\, W_i = 0\}} Y_i, \tag{2.3}$$

where $n_x = |\{i : X_i = x\}|$ and $n_{xw} = |\{i : X_i = x,\, W_i = w\}|$. How good is this estimator? Intuitively, we have needed to estimate $|\mathcal{X}| = p$ "parameters" so we might expect the variance to scale linearly with p?

To study this estimator it is helpful to write it as follows. First, for any group with covariate x, define $e(x)$ as the probability of getting treated in that group, $e(x) = \mathbb{P}\left[W_i = 1 \,\middle|\, X_i = x\right]$, and note that

$$\sqrt{n_x}\left(\hat{\tau}(x) - \tau(x)\right) \Rightarrow \mathcal{N}\left(0,\, \frac{\text{Var}\left[Y_i(0) \,\middle|\, X_i = x\right]}{1 - e(x)} + \frac{\text{Var}\left[Y_i(1) \,\middle|\, X_i = x\right]}{e(x)}\right).$$

Furthermore, under the simplifying assumption that $\text{Var}\left[Y(w) \,\middle|\, X = x\right] = \sigma^2(x)$ does not depend on w, we get

$$\sqrt{n_x}\left(\hat{\tau}(x) - \tau(x)\right) \Rightarrow \mathcal{N}\left(0,\, \frac{\sigma^2(x)}{e(x)(1 - e(x))}\right). \tag{2.4}$$

Next, for the aggregated estimator, defining $\hat{\pi}(x) = n_x/n$ as the fraction of observations with $X_i = x$ and $\pi(x) = \mathbb{P}\left[X_i = x\right]$ as its expectation, we have

$$\begin{aligned}\hat{\tau}_{AGG} = \sum_{x\in\mathcal{X}} \hat{\pi}(x)\hat{\tau}(x) = \underbrace{\sum_{x\in\mathcal{X}} \pi(x)\tau(x)}_{=\tau} + \underbrace{\sum_{x\in\mathcal{X}} \pi(x)\left(\hat{\tau}(x) - \tau(x)\right)}_{\approx\mathcal{N}\left(0,\, \sum_{x\in\mathcal{X}} \pi^2(x)\,\text{Var}[\hat{\tau}(x)]\right)} \\ + \underbrace{\sum_{x\in\mathcal{X}} \left(\hat{\pi}(x) - \pi(x)\right)\tau(x)}_{\approx\mathcal{N}(0,\, n^{-1}\,\text{Var}[\tau(X_i)])} + \underbrace{\sum_{x\in\mathcal{X}} \left(\hat{\pi}(x) - \pi(x)\right)\left(\hat{\tau}(x) - \tau(x)\right)}_{=\mathcal{O}_P(1/n)}.\end{aligned}$$

Putting the pieces together, we get $\sqrt{n}\left(\hat{\tau}_{AGG} - \tau\right) \Rightarrow \mathcal{N}\left(0,\, V_{AGG}\right)$

$$\begin{aligned}V_{AGG} &= \text{Var}\left[\tau(X_i)\right] + \sum_{x\in\mathcal{X}} \pi^2(x)\frac{1}{\pi(x)}\frac{\sigma^2(x)}{e(x)(1 - e(x))} \\ &= \text{Var}\left[\tau(X_i)\right] + \mathbb{E}\left[\frac{\sigma^2(X_i)}{e(X_i)(1 - e(X_i))}\right].\end{aligned} \tag{2.5}$$

Note that this does not depend on $|\mathcal{X}| = p$, the number of groups(!)

Continuous X and the propensity score Above, we considered a setting where $\mathcal{X}$ is discrete with a finite number levels, and treatment W_i is as good as random conditionally on $X_i = x$ as in (2.1). In this case, we found that we can still accurately estimate the ATE by aggregating group-wise treatment

effect estimates, and that the exact number of groups $|\mathcal{X}| = p$ does not affect the accuracy of inference. However, if $\mathcal{X}$ is continuous (or the cardinality of $\mathcal{X}$ is very large), this result does not apply directly—because we won't be able to get enough samples for each possible value of $x \in \mathcal{X}$ to be able to define $\hat{\tau}(x)$ as in (2.3).

In order to generalize our analysis beyond the discrete-X case, we'll need to move beyond literally trying to estimate $\tau(x)$ for each value of x by simple averaging, and use a more indirect argument instead. To this end, we first need to generalize the "RCT in each group" assumption. Formally, we just write the same thing,

$$\{Y_i(0),\, Y_i(1)\} \perp\!\!\!\perp W_i \,\big|\, X_i, \tag{2.6}$$

although now X_i may be an arbitrary random variable, and interpretation of this statement may require more care. Qualitatively, one way to think about (2.6) is that we have measured enough covariates to capture any dependence between W_i and the potential outcomes and so, given X_i, W_i cannot "peek" at the $\{Y_i(0),\, Y_i(1)\}$. We call this assumption **unconfoundedness**.

The assumption (2.6) may seem like a difficult assumption to use in practice, since it involves conditioning on a continuous random variable. However, as shown by Rosenbaum and Rubin (1983), this assumption can be made considerably more tractable by considering the **propensity score**[1]

$$e(x) = \mathbb{P}\left[W_i = 1 \,\middle|\, X_i = x\right]. \tag{2.7}$$

Statistically, a key property of the propensity score is that it is a balancing score: If (2.6) holds, then in fact

$$\{Y_i(0),\, Y_i(1)\} \perp\!\!\!\perp W_i \,\big|\, e(X_i), \tag{2.8}$$

i.e., it actually suffices to control for $e(X)$ rather than X to remove biases associated with a non-random treatment assignment. We can verify this claim as follows:

$$\begin{aligned}
&\mathbb{P}\left[W_i = w \,\middle|\, \{Y_i(0),\, Y_i(1)\}, e(X_i)\right] \\
&\quad = \int_{\mathcal{X}} \mathbb{P}\left[W_i = w \,\middle|\, \{Y_i(0),\, Y_i(1)\}, X_i = x\right] \mathbb{P}\left[X_i = x \,\middle|\, e(X_i)\right]\; dx \\
&\quad = \int_{\mathcal{X}} \mathbb{P}\left[W_i = w \,\middle|\, X_i = x\right] \mathbb{P}\left[X_i = x \,\middle|\, e(X_i)\right]\; dx \qquad \text{(unconf.)} \\
&\quad = e(X_i)\mathbf{1}_{w=1} + (1 - e(X_i))\mathbf{1}_{w=0}.
\end{aligned}$$

[1] When X is continuous, the propensity score $e(x)$ has exactly the same meaning as when X is discrete; however, we can no longer trivially estimate it via $\hat{e}(x) = n_{x1}/n_x$ in this case.

The implication of (2.8) is that if we can partition our observations into groups with (almost) constant values of the propensity score $e(x)$, then we can consistently estimate the average treatment effect via variants of $\hat{\tau}_{AGG}$.

Propensity stratification One instantiation of this idea is propensity stratification, which proceeds as follows. First obtain an estimate $\hat{e}(x)$ of the propensity score via non-parametric regression, and choose a number of strata J. Then:

1. Sort the observations according to their propensity scores, such that

$$\hat{e}\left(X_{i_1}\right) \leq \hat{e}\left(X_{i_2}\right) \leq \ldots \leq \hat{e}\left(X_{i_n}\right). \tag{2.9}$$

2. Split the sample into J evenly size strata using the sorted propensity score and, in each stratum $j = 1, ..., J$, compute the simple difference-in-means treatment effect estimator for the stratum:

$$\hat{\tau}_j = \frac{\sum_{j=\lfloor (j-1)n/J \rfloor + 1}^{\lfloor jn/J \rfloor} W_i Y_i}{\sum_{j=\lfloor (j-1)n/J \rfloor + 1}^{\lfloor jn/J \rfloor} W_i} - \frac{\sum_{j=\lfloor (j-1)n/J \rfloor + 1}^{\lfloor jn/J \rfloor} (1 - W_i) Y_i}{\sum_{j=\lfloor (j-1)n/J \rfloor + 1}^{\lfloor jn/J \rfloor} (1 - W_i)}. \tag{2.10}$$

3. Estimate the average treatment by applying the idea of (2.3) across strata:

$$\hat{\tau}_{STRAT} = \frac{1}{J} \sum_{j=1}^{J} \hat{\tau}_j. \tag{2.11}$$

The arguments described above immediately imply that, thanks to (2.8), $\hat{\tau}_{STRAT}$ is consistent for τ whenever $\hat{e}(x)$ is uniformly consistent for $e(x)$ and the number of strata J grows appropriately with n.

Inverse-propensity weighting Another, algorithmically simpler way of exploiting unconfoundedness is via inverse-propensity weighting: As before, we first estimate $\hat{e}(x)$ via non-parametric regression, and then set

$$\hat{\tau}_{IPW} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{W_i Y_i}{\hat{e}(X_i)} - \frac{(1 - W_i) Y_i}{1 - \hat{e}(X_i)} \right). \tag{2.12}$$

The simplest way to analyze it is by comparing it to an oracle that actually knows the propensity score:

$$\hat{\tau}^*_{IPW} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{W_i Y_i}{e(X_i)} - \frac{(1 - W_i) Y_i}{1 - e(X_i)} \right). \tag{2.13}$$

Suppose that we have **overlap**, i.e., that

$$\eta \leq e(x) \leq 1 - \eta \text{ for all } x \in \mathcal{X}. \tag{2.14}$$

Suppose moreover that $|Y_i| \leq M$, and that we know that $\sup_{x \in \mathcal{X}} |e(x) - \hat{e}(x)| = \mathcal{O}_P(a_n) \to 0$. Then, we can check that

$$|\hat{\tau}_{IPW} - \hat{\tau}^*_{IPW}| = \mathcal{O}_P\left(\frac{a_n M}{\eta}\right), \tag{2.15}$$

and so if $\hat{\tau}^*_{IPW}$ is consistent, then so is $\hat{\tau}_{IPW}$.

It thus remains to analyze the behavior of the oracle IPW estimator $\hat{\tau}^*_{IPW}$. First, we note that

$$\begin{aligned}
\mathbb{E}\left[\hat{\tau}^*_{IPW}\right] &= \mathbb{E}\left[\frac{W_i Y_i}{e(X_i)} - \frac{(1-W_i)Y_i}{1-e(X_i)}\right] && \text{(IID)} \\
&= \mathbb{E}\left[\frac{W_i Y_i(1)}{e(X_i)} - \frac{(1-W_i)Y_i(0)}{1-e(X_i)}\right] && \text{(SUTVA)} \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{W_i Y_i(1)}{e(X_i)} \,\middle|\, e(X_i)\right] - \mathbb{E}\left[\frac{(1-W_i)Y_i(0)}{1-e(X_i)} \,\middle|\, e(X_i)\right]\right] \\
&= \mathbb{E}\left[Y_i(1) - Y_i(0)\right] && \text{(unconf.)},
\end{aligned}$$

meaning that the oracle estimator is unbiased τ. Meanwhile, under overlap (2.14), we immediately see that $\hat{\tau}^*_{IPW}$ concentrates at $1/\sqrt{n}$-rates; and thus $\hat{\tau}^*_{IPW}$ is consistent for τ.

The variance of oracle IPW Studying the accuracy of IPW in a way that properly accounts for the behavior of the estimated propensity scores $\hat{e}(x)$ is somewhat delicate, and intricately depends on the choice of estimator $\hat{e}(x)$. Thus, let's start by considering the accuracy of the oracle $\hat{\tau}^*_{IPW}$. We already know that it is unbiased, and so we only need to express its variance. To do so, it is helpful to expand out (without loss of generality),[2]

$$\begin{aligned}
Y_i(0) &= c(X_i) - (1 - e(X_i))\tau(X_i) + \varepsilon_i(0), \quad \mathbb{E}\left[\varepsilon_i(0) \,\middle|\, X_i\right] = 0 \\
Y_i(1) &= c(X_i) + e(X_i)\tau(X_i) + \varepsilon_i(1), \quad \mathbb{E}\left[\varepsilon_i(1) \,\middle|\, X_i\right] = 0,
\end{aligned} \tag{2.16}$$

and assume for simplicity that $\text{Var}\left[\varepsilon_i(w) \,\middle|\, X_i = x\right] = \sigma^2(x)$ does not depend on w. Then, we can verify that (on the second line, the fact that the variances

[2]In particular, note that $\mathbb{E}\left[Y_i(1) - Y_i(0) \,\middle|\, X_i = x\right] = \tau(x)$. Here, the function $c(x)$ is simply chosen such as to make the decomposition (2.16) work.

separate is non-trivial, and is a result of how we defined $c(\cdot)$)

$$\begin{aligned} n \operatorname{Var}\left[\hat{\tau}^*_{IPW}\right] &= \operatorname{Var}\left[\frac{W_i Y_i}{e(X_i)} - \frac{(1-W_i)Y_i}{1-e(X_i)}\right] \\ &= \operatorname{Var}\left[\frac{W_i c(X_i)}{e(X_i)} - \frac{(1-W_i)c(X_i)}{1-e(X_i)}\right] + \operatorname{Var}\left[\tau(X_i)\right] \\ &\quad + \operatorname{Var}\left[\frac{W_i \varepsilon_i}{e(X_i)} - \frac{(1-W_i)\varepsilon_i}{1-e(X_i)}\right] \\ &= \mathbb{E}\left[\frac{c^2(X_i)}{e(X_i)(1-e(X_i))}\right] + \operatorname{Var}\left[\tau(X_i)\right] + \mathbb{E}\left[\frac{\sigma^2(X_i)}{e(X_i)(1-e(X_i))}\right]. \end{aligned}$$

Pulling everything together, we see that

$$\begin{aligned} &\sqrt{n}\left(\hat{\tau}^*_{IPW} - \tau\right) \Rightarrow \mathcal{N}\left(0,\, V_{IPW^*}\right), \\ &V_{IPW^*} = \mathbb{E}\left[\frac{c^2(X_i)}{e(X_i)(1-e(X_i))}\right] \\ &\qquad\qquad + \operatorname{Var}\left[\tau(X_i)\right] + \mathbb{E}\left[\frac{\sigma^2(X_i)}{e(X_i)(1-e(X_i))}\right]. \end{aligned} \tag{2.17}$$

How accurate is oracle IPW? To gain a better understanding of how good the accuracy is, it is helpful to re-visit the setting of the beginning of this lecture where $\mathcal{X}$ is discrete. In this setting, nothing's stopping us from using IPW; but we now can also use our group-wise aggregated estimator $\hat{\tau}_{AGG}$ from (2.3) as a point of comparison. And, in doing so, we see that the performance of the oracle IPW estimator is somewhat disappointing. Despite having access to the true propensity score $e(x)$, it always under-performs $\hat{\tau}_{AGG}$: Both estimators are asymptotically centered normal, but from (2.5) and (2.17) we see that

$$V_{IPW*} = V_{AGG} + \mathbb{E}\left[\frac{c^2(X_i)}{e(X_i)(1-e(X_i))}\right]. \tag{2.18}$$

Thus, unless $c(x)$ as defined via (2.16) is zero everywhere, $\hat{\tau}^*_{IPW}$ has a strictly worse asymptotic variance than $\hat{\tau}_{AGG}$.

Perhaps even more surprisingly, we note that $\hat{\tau}_{AGG}$ can actually be understood as an IPW estimator with a specific choice of estimated propensity score $\hat{e}(x)$:

$$\begin{aligned} \hat{\tau}_{AGG} &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{W_i Y_i}{\hat{e}(X_i)} - \frac{(1-W_i)\,Y_i}{1-\hat{e}(X_i)}\right), \quad \hat{e}(x) = \frac{n_{x1}}{n_1}, \\ \hat{\tau}^*_{IPW} &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{W_i Y_i}{e(X_i)} - \frac{(1-W_i)\,Y_i}{1-e(X_i)}\right). \end{aligned} \tag{2.19}$$

Thus, the "feasible" IPW estimator $\hat{\tau}_{AGG}$ is actually better than the "oracle" IPW estimator. At a high level, the reason this phenomenon occurs is that the estimated propensity score corrects for local variability in the sampling distribution of the W_i (i.e., it accounts for the number of units that were *actually* treated in each group).

Comparison with linear modeling One can contrast this approach to a "classical" approach to controlling for covariates based on parametric modeling. In such a classical analysis, one might estimate the effect of a non-randomized treatment W_i by writing down a linear regression model

$$Y_i \sim X_i\beta + W_i\tau, \tag{2.20}$$

and then estimating the model by OLS. One might then argue that τ is the effect of W_i while "controlling" for X_i.

The approach following (2.20) is potentially acceptable if one knows the linear model to be well specified, or is willing to settle for a more heuristic analysis. However, one should note that the standard of rigor underlying such linear modeling vs. the methods discussed today is quite different. As discussed today, IPW is consistent under the substantively meaningful assumption (2.6), whereby treatment assigned emulates random treatment assignment once we control for X_i. On the other hand, the linear modeling approach is entirely dependent on well-specification of (2.20); and in case of model misspecification, there's no reason to except that its $\hat{\tau}$ estimate will converge to anything that can be interpreted as a causal effect.

Recap Today, we discussed estimation of the average treatment effect under unconfoundedness, i.e., under the assumption that we observe a set of covariates X_i such that treatment is as good as random after we control for X_i in the sense of (2.6), and showed that estimators based on the propensity score achieve non-parametric consistency.

We found that IPW is a simple estimator that easily enables us to exploit unconfoundedness, and with true propensity score $se(x)$ it is unbiased. However, we also found that a variant of IPW with estimated propensity scores can, in some cases, outperform the oracle IPW estimator. This provides evidence that IPW is not "optimal," and does not fully capture the complexity of the problem of average treatment effect estimation under unconfoundedness. In the following lecture, we'll discuss alternatives to IPW with better asymptotic properties.

Bibliographic notes The central role of the propensity score in estimating causal effects was first emphasized by Rosenbaum and Rubin [1983], while associated methods for estimation such as propensity stratification are discussed in Rosenbaum and Rubin [1984]. Hirano, Imbens, and Ridder [2003] provide a detailed discussion of the asymptotics of IPW-style estimators; and in particular they discuss conditions under which IPW with non-parametrically estimated propensity scores can outperform oracle IPW.

Another popular way of leveraging the propensity score in practice is propensity matching, i.e., estimating treatment effects by comparing pairs of units with similar values of $\hat{e}(X_i)$. For a some recent discussions of matching in causal inference, see Abadie and Imbens [2006, 2016], Diamond and Sekhon [2013], Zubizarreta [2012], and references therein.

Imbens [2004] provides a general overview of methods for treatment effect estimation under unconfoundedness, including a discussion of alternative estimands to the average treatment effect, such as the average treatment effect on the treated.

Lecture 3
Efficient Treatment Effect Estimation via Augmented IPW

Inverse-propensity weighting (IPW) is a simple and transparent approach to average treatment effect estimation under unconfoundedness. However, as seen in the previous lecture, the large-sample properties of IPW are not particularly good in general. For example, in the case where the covariates $X_i \in \mathcal{X}$ are discrete, we found that IPW underperforms a baseline that estimates separate treatment effects for each value of $x \in \mathcal{X}$ and then aggregates them. The goal of this lecture is to get beyond the limitations of IPW, and to discuss a general recipe for building asymptotically optimal treatment effect estimators under unconfoundedness.

Statistical setting We observe data $(X_i,\, Y_i,\, W_i,\,) \in \mathcal{X} \times \mathbb{R} \times \{0,\, 1\}$ according to the potential outcomes model, such that there are potential outcomes $\{Y_i(0),\, Y_i(1)\}$ for which $Y_i = Y_i(W_i)$ (SUTVA). We are not necessarily in a randomized controlled trial; however, we assume unconfoundedness, i.e., that treatment assignment is as good as random conditionally on the features X_i:

$$\{Y_i(0),\, Y_i(1)\} \perp\!\!\!\perp W_i \,\big|\, X_i. \tag{3.1}$$

We seek to estimate the average treatment effect $\tau = \mathbb{E}\left[Y_i(1) - Y_i(0)\right]$. Throughout, we write $\sigma_w^2(x) = \mathrm{Var}\left[Y_i(w) \,\middle|\, X_i = x\right]$.

Two characterizations of the ATE Last time, we saw that the ATE can be characterized in terms of the propensity score $e(x) = \mathbb{P}\left[W_i = 1 \,\middle|\, X_i = x\right]$:

$$\tau = \mathbb{E}\left[\hat{\tau}^*_{IPW}\right], \quad \hat{\tau}^*_{IPW} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{W_i Y_i}{e(X_i)} - \frac{(1-W_i)\,Y_i}{1-e(X_i)}\right). \tag{3.2}$$

However, τ can also be characterized in terms of the conditional response surfaces $\mu_{(w)}(x) = \mathbb{E}\left[Y_i(w) \,\middle|\, X_i = x\right]$. Under unconfoundedness (3.1),

$$\begin{aligned}
\tau(x) &:= \mathbb{E}\left[Y_i(1) - Y_i(0) \,\middle|\, X_i = x\right] \\
&= \mathbb{E}\left[Y_i(1) \,\middle|\, X_i = x\right] - \mathbb{E}\left[Y_i(0) \,\middle|\, X_i = x\right] \\
&= \mathbb{E}\left[Y_i(1) \,\middle|\, X_i = x,\, W_i = 1\right] - \mathbb{E}\left[Y_i(0) \,\middle|\, X_i = x,\, W_i = 0\right] \quad \text{(unconf)} \\
&= \mathbb{E}\left[Y_i \,\middle|\, X_i = x,\, W_i = 1\right] - \mathbb{E}\left[Y_i \,\middle|\, X_i = x,\, W_i = 0\right] \quad \text{(SUTVA)} \\
&= \mu_{(1)}(x) - \mu_{(0)}(x),
\end{aligned}$$

and so $\tau = \mathbb{E}\left[\mu_{(1)}(x) - \mu_{(0)}(x)\right]$. Thus we could also derive a consistent (but not necessarily optimal) estimator for τ by first estimating $\mu_{(0)}(x)$ and $\mu_{(1)}(x)$ non-parametrically, and then using $\hat{\tau}_{REG} = n^{-1}\sum_{i=1}^{n}(\hat{\mu}_{(1)}(X_i) - \hat{\mu}_{(0)}(X_i))$.

Augmented IPW Given that the average treatment effect can be estimated in two different ways, i.e., by first non-parametrically estimating $e(x)$ or by first estimating $\mu_{(0)}(x)$ and $\mu_{(1)}(x)$, it is natural to ask whether it is possible to combine both strategies. This turns out to be a very good idea, and yields the augmented IPW (AIPW) estimator of Robins, Rotnitzky, and Zhao [1994]:

$$\begin{aligned}
\hat{\tau}_{AIPW} = \frac{1}{n}\sum_{i=1}^{n}\Bigg(&\hat{\mu}_{(1)}(X_i) - \hat{\mu}_{(0)}(X_i) \\
&+ W_i\,\frac{Y_i - \hat{\mu}_{(1)}(X_i)}{\hat{e}(X_i)} - (1 - W_i)\,\frac{Y_i - \hat{\mu}_{(0)}(X_i)}{1 - \hat{e}(X_i)}\Bigg).
\end{aligned} \tag{3.3}$$

Qualitatively, AIPW can be seen as first making a best effort attempt at τ by estimating $\mu_{(0)}(x)$ and $\mu_{(1)}(x)$; then, it deals with any biases of the $\hat{\mu}_{(w)}(x)$ by applying IPW to the regression residuals.

Double robustness AIPW has many good statistical properties. One of its properties that is easiest to explain is "double robustness": AIPW is consistent if either the $\hat{\mu}_{(w)}(x)$ are consistent or $\hat{e}(x)$ is consistent. To see this, first consider the case where $\hat{\mu}_{(w)}(x)$ is consistent, i.e., $\hat{\mu}_{(w)}(x) \approx \mu_{(w)}(x)$. Then,

$$\begin{aligned}
\hat{\tau}_{AIPW} = &\underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}_{(1)}(X_i) - \hat{\mu}_{(0)}(X_i)\right)}_{\text{a consistent treatment effect estimator}} \\
&+ \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{W_i}{\hat{e}(X_i)}\left(Y_i - \hat{\mu}_{(1)}(X_i)\right) - \frac{1 - W_i}{1 - \hat{e}(X_i)}\left(Y_i - \hat{\mu}_{(0)}(X_i)\right)\right)}_{\approx \text{ mean-zero noise}},
\end{aligned}$$

because $\mathbb{E}\left[Y_i - \hat{\mu}_{(W_i)}(X_i) \,\middle|\, X_i, W_i\right] \approx 0$, and so the "garbage" propensity score weight $1/\hat{e}(X_i)$, resp. $1/(1-\hat{e}(X_i))$ is multiplied by mean-zero noise that makes it go away. Thus $\hat{\tau}_{AIPW}$ is consistent. Second, suppose that $\hat{e}(x)$ is consistent, i.e., $\hat{e}(x) \approx e(x)$. Then,

$$\hat{\tau}_{AIPW} = \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{W_i Y_i}{\hat{e}(X_i)} - \frac{(1-W_i)Y_i}{1-\hat{e}(X_i)}\right)}_{\text{the IPW estimator}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}_{(1)}(X_i)\left(1-\frac{W_i}{\hat{e}(X_i)}\right) - \hat{\mu}_{(0)}(X_i)\left(1-\frac{1-W_i}{1-\hat{e}(X_i)}\right)\right)}_{\approx \text{ mean-zero noise}},$$

because $\mathbb{E}\left[1 - W_i/\hat{e}(X_i) \,\middle|\, X_i\right] \approx 0$, and so the "garbage" regression adjustments $\hat{\mu}_{(w)}(X_i)$ is multiplied by mean-zero noise that makes it go away. Thus $\hat{\tau}_{AIPW}$ is consistent.

The double robustness of AIPW is well known; in fact AIPW is sometimes referred to as *the* doubly robust estimator—although there are many others. My own view is that while double robustness is a nice property to have, its importance should not be overstated. In a modern statistical setting, we should be using appropriate non-parametric estimators for both $\mu_{(w)}(x)$ and $e(x)$ such that both are consistent; in which case the double robustness statement doesn't buy us much, while the conclusion of the double robustness argument (namely consistency of $\hat{\tau}_{AIPW}$) is rather weak.

Semiparametric efficiency The more important property of AIPW is that it is asymptotically optimal among all non-parametric estimators in a strong sense. Provided we estimate $\mu_{(w)}(x)$ and $e(x)$ in a reasonably accurate way (and we'll discuss specific conditions under which this holds in just a minute), one can show that $\hat{\tau}_{AIPW}$ is to first order equivalent to the oracle AIPW estimator

$$\hat{\tau}^*_{AIPW} = \frac{1}{n}\sum_{i=1}^{n}\Bigg(\mu_{(1)}(X_i) - \mu_{(0)}(X_i) + W_i\,\frac{Y_i - \mu_{(1)}(X_i)}{e(X_i)} - (1-W_i)\,\frac{Y_i - \mu_{(0)}(X_i)}{1-e(X_i)}\Bigg), \tag{3.4}$$

meaning that

$$\sqrt{n}\,(\hat{\tau}_{AIPW} - \hat{\tau}^*_{AIPW}) \to_p 0. \tag{3.5}$$

Now, $\hat{\tau}^*_{AIPW}$ is just an IID average, so we immediately see that[1]

$$\begin{aligned}
&\sqrt{n}\left(\hat{\tau}^*_{AIPW} - \tau\right) \Rightarrow \mathcal{N}\left(0,\, V^*\right), \\
&V^* = \mathrm{Var}\left[\tau(X_i)\right] + \mathbb{E}\left[\frac{\sigma_0^2(X_i)}{1-e(X_i)}\right] + \mathbb{E}\left[\frac{\sigma_1^2(X_i)}{e(X_i)}\right],
\end{aligned} \tag{3.6}$$

and so whenever (3.5) holds $\hat{\tau}_{AIPW}$ also satisfies a CLT as in (3.6). Furthermore, it turns out that the behavior (3.6) is asymptotically optimal, in the sense that no "regular" estimator of τ can improve on the behavior in (3.6).[2] This result is a Cramer-Rao type bound for non-parametric average treatment effect estimation.[3]

AIPW and cross-fitting When choosing which treatment effect estimator to use in practice, we want to attain performance as in (3.6) and so need to make sure that (3.5) holds. To this end, consider the following minor modification of AIPW using cross-fitting. At a high level, cross-fitting uses cross-fold estimation to avoid bias due to overfitting; the reason why this works is exactly the same as why we want to use cross-validation when estimating the predictive accuracy of an estimator.

Cross-fitting first splits the data (at random) into two halves $\mathcal{I}_1$ and $\mathcal{I}_2$, and then uses an estimator[4]

$$\begin{aligned}
\hat{\tau}_{AIPW} =& \frac{|\mathcal{I}_1|}{n}\,\hat{\tau}^{\mathcal{I}_1} + \frac{|\mathcal{I}_2|}{n}\,\hat{\tau}^{\mathcal{I}_2}, \qquad \hat{\tau}^{\mathcal{I}_1} = \frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\Bigg(\hat{\mu}^{\mathcal{I}_2}_{(1)}(X_i) - \hat{\mu}^{\mathcal{I}_2}_{(0)}(X_i) \\
&+ W_i\,\frac{Y_i - \hat{\mu}^{\mathcal{I}_2}_{(1)}(X_i)}{\hat{e}^{\mathcal{I}_2}(X_i)} - (1-W_i)\,\frac{Y_i - \hat{\mu}^{\mathcal{I}_2}_{(0)}(X_i)}{1-\hat{e}^{\mathcal{I}_2}(X_i)}\Bigg),
\end{aligned} \tag{3.7}$$

where the $\hat{\mu}^{\mathcal{I}_2}_{(w)}(\cdot)$ and $\hat{e}^{\mathcal{I}_2}(\cdot)$ are estimates of $\mu_{(w)}(\cdot)$ and $e(\cdot)$ obtained using only the half-sample $\mathcal{I}_2$, and $\hat{\tau}^{\mathcal{I}_2}$ is defined analogously (with the roles of $\mathcal{I}_1$

[1] To see why $\hat{\tau}^*$ has variance V^*/n, note that we can decompose its summands into 3 uncorrelated parts: $\mu_{(1)}(X_i) - \mu_{(0)}(X_i)$, $W_i\left(Y_i - \mu_{(1)}(X_i)\right)/e(X_i)$, and $(1 - W_i)\left(Y_i - \mu_{(0)}(X_i)\right)/\left(1 - e(X_i)\right)$.

[2] Interestingly, note that the estimator $\hat{\tau}_{AGG}$ discussed in the last class for the case where $\mathcal{X}$ is discrete also had asymptotic variance V^*, and is thus semiparametrically efficient. There is a large taxonomy of different ATE estimators under unconfoundedness; but the expectation is that all the good ones should attain efficiency.

[3] A discussion of why the behavior (3.6) is optimal is beyond the scope of this class and instead belongs in a class on theoretical statistic and/or semiparametrics; however, for those of you who are curious to see an argument, Hahn [1998] is a good place to start.

[4] In subsequent lectures, whenever I'll talk about AIPW, I'll implicitly assume we're using cross-fitting unless specified otherwise.

and $\mathcal{I}_2$ swapped). In other words, $\hat{\tau}^{\mathcal{I}_1}$ is a treatment effect estimator on $\mathcal{I}_1$ that uses $\mathcal{I}_2$ to estimate its nuisance components, and vice-versa.

This cross-estimation construction allows us to, asymptotically, ignore the idiosyncrasies of the specific machine learning adjustment we chose to use, and to simply rely on the following high-level conditions:

1. **Overlap:** The true propensity score is bounded away from 0 and 1, such that $\eta < e(x) < 1 - \eta$ for all $x \in \mathcal{X}$.

2. **Consistency:** All machine learning adjustments are sup-norm consistent,
$$\sup_{x \in \mathcal{X}} \left| \hat{\mu}_{(w)}^{\mathcal{I}_2}(x) - \mu_{(w)}(x) \right|, \quad \sup_{x \in \mathcal{X}} \left| \hat{e}^{\mathcal{I}_2}(x) - e(x) \right| \to_p 0.$$

3. **Risk decay:** The product of the errors for the outcome and propensity models decays as
$$\mathbb{E}\left[\left(\hat{\mu}_{(w)}^{\mathcal{I}_2}(X_i) - \mu_{(w)}(X_i) \right)^2 \right] \mathbb{E}\left[\left(\hat{e}^{\mathcal{I}_2}(X_i) - e(X_i) \right)^2 \right] = o\left(\frac{1}{n} \right), \quad (3.8)$$
where the randomness above is taken over both the training of $\hat{\mu}_{(w)}$ and $\hat{e}$ and the test example X. Note that if $\hat{\mu}_{(w)}$ and $\hat{e}$ both attained the parametric "$\sqrt{n}$-consistent" rate, then the error product would be bounded as $\mathcal{O}(1/n^2)$. A simple way to satisfy this condition is to have all regression adjustments be $o(n^{-1/4})$ consistent in root-mean squared error (RMSE).

Note that none of these conditions depend on the internal structure of the machine learning method used. Moreover, (3) depends on the mean-squared error of the risk adjustments, and so justifies tuning the $\hat{\mu}_{(w)}$ and $\hat{e}$ estimates via cross-validation.

Given these assumptions, we characterize the cross-fitting estimator (3.7) by coupling it with the oracle efficient score estimator (3.4), i.e.,
$$\sqrt{n}\left(\hat{\tau}_{AIPW} - \hat{\tau}^* \right) \to_p 0. \quad (3.9)$$
To do so, we first note that we can write
$$\hat{\tau}^* = \frac{|\mathcal{I}_1|}{n} \hat{\tau}^{\mathcal{I}_1,*} + \frac{|\mathcal{I}_2|}{n} \hat{\tau}^{\mathcal{I}_2,*}$$
analogously to (3.7) (because $\hat{\tau}^*$ uses oracle nuisance components, the cross-fitting construction doesn't change anything for it). Moreover, we can decompose $\hat{\tau}^{\mathcal{I}_1}$ itself as
$$\begin{aligned} \hat{\tau}^{\mathcal{I}_1} &= \hat{\mu}_{(1)}^{\mathcal{I}_1} - \hat{\mu}_{(0)}^{\mathcal{I}_1}, \\ \hat{\mu}_{(1)}^{\mathcal{I}_1} &= \frac{1}{|\mathcal{I}_1|} \sum_{i \in \mathcal{I}_1} \left(\hat{\mu}_{(1)}^{\mathcal{I}_2}(X_i) + W_i \frac{Y_i - \hat{\mu}_{(1)}^{\mathcal{I}_2}(X_i)}{\hat{e}^{\mathcal{I}_2}(X_i)} \right), \end{aligned} \quad (3.10)$$

etc., and define $\hat{\mu}^{\mathcal{I}_1,*}_{(0)}$ and $\hat{\mu}^{\mathcal{I}_1,*}_{(1)}$ analogously. Given this buildup, in order to verify (3.9), it suffices to show that

$$\sqrt{n}\left(\hat{\mu}^{\mathcal{I}_1}_{(1)} - \hat{\mu}^{\mathcal{I}_1,*}_{(1)}\right) \to_p 0, \tag{3.11}$$

etc., across folds and treatment statuses.

We now study the term in (3.11) by decomposing it as follows:

$$\begin{aligned}
&\hat{\mu}^{\mathcal{I}_1}_{(1)} - \hat{\mu}^{\mathcal{I}_1,*}_{(1)} \\
&= \frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\left(\hat{\mu}^{\mathcal{I}_2}_{(1)}(X_i) + W_i\frac{Y_i - \hat{\mu}^{\mathcal{I}_2}_{(1)}(X_i)}{\hat{e}^{\mathcal{I}_2}(X_i)} - \mu_{(1)}(X_i) - W_i\frac{Y_i - \mu_{(1)}(X_i)}{e(X_i)}\right) \\
&= \frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\left(\left(\hat{\mu}^{\mathcal{I}_2}_{(1)}(X_i) - \mu_{(1)}(X_i)\right)\left(1 - \frac{W_i}{e(X_i)}\right)\right) \\
&\quad + \frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1} W_i\left(\left(Y_i - \mu_{(1)}(X_i)\right)\left(\frac{1}{\hat{e}^{\mathcal{I}_2}(X_i)} - \frac{1}{e(X_i)}\right)\right) \\
&\quad - \frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1} W_i\left(\left(\hat{\mu}^{\mathcal{I}_2}_{(1)}(X_i) - \mu_{(1)}(X_i)\right)\left(\frac{1}{\hat{e}^{\mathcal{I}_2}(X_i)} - \frac{1}{e(X_i)}\right)\right)
\end{aligned}$$

Now, we can verify that these are small for different reasons. For the first term, we intricately use the fact that, thanks to our double machine learning construction, $\hat{\mu}^{\mathcal{I}_2}_{(w)}$ can effectively be treated as deterministic. Thus after conditioning on $\mathcal{I}_2$, the summands used to build this term become mean-zero and independent (2nd and 3rd equalities below)

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\left(\left(\hat{\mu}^{\mathcal{I}_2}_{(1)}(X_i) - \mu_{(1)}(X_i)\right)\left(1 - \frac{W_i}{e(X_i)}\right)\right)\right)^2\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left(\frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\left(\left(\hat{\mu}^{\mathcal{I}_2}_{(1)}(X_i) - \mu_{(1)}(X_i)\right)\left(1 - \frac{W_i}{e(X_i)}\right)\right)\right)^2 \,\middle|\, \mathcal{I}_2\right]\right] \\
&= \mathbb{E}\left[\mathrm{Var}\left[\frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\left(\left(\hat{\mu}^{\mathcal{I}_2}_{(1)}(X_i) - \mu_{(1)}(X_i)\right)\left(1 - \frac{W_i}{e(X_i)}\right)\right) \,\middle|\, \mathcal{I}_2\right]\right] \\
&= \frac{1}{|\mathcal{I}_1|}\mathbb{E}\left[\mathrm{Var}\left[\left(\hat{\mu}^{\mathcal{I}_2}_{(1)}(X_i) - \mu_{(1)}(X_i)\right)\left(1 - \frac{W_i}{e(X_i)}\right) \,\middle|\, \mathcal{I}_2\right]\right] \\
&= \frac{1}{|\mathcal{I}_1|}\mathbb{E}\left[\mathbb{E}\left[\left(\hat{\mu}^{\mathcal{I}_2}_{(1)}(X_i) - \mu_{(1)}(X_i)\right)^2\left(\frac{1}{e(X_i)} - 1\right) \,\middle|\, \mathcal{I}_2\right]\right] \\
&\leq \frac{1}{\eta\,|\mathcal{I}_1|}\mathbb{E}\left[\left(\hat{\mu}^{\mathcal{I}_2}_{(1)}(X_i) - \mu_{(1)}(X_i)\right)^2\right] = \frac{o_P(1)}{n}
\end{aligned}$$

by consistency (2), because $\mathcal{I}_1 \sim n/2$. The key step in this argument was the 3rd equality: Because the summands become independent and mean-zero after conditioning, we "earn" a factor $1/|\mathcal{I}_1|$ due to concentration of iid sums. The second summand in our decomposition here can also be bounded similarly (thanks to overlap). Finally, for the last summand, we simply use Cauchy-Schwarz:

$$\begin{aligned}
&\frac{1}{|\mathcal{I}_1|} \sum_{\{i: i \in \mathcal{I}_1,\, W_i = 1\}} \left(\left(\hat{\mu}_{(1)}^{\mathcal{I}_2}(X_i) - \mu_{(1)}(X_i) \right) \left(\frac{1}{\hat{e}^{\mathcal{I}_2}(X_i)} - \frac{1}{e(X_i)} \right) \right) \\
&\quad \leq \sqrt{\frac{1}{|\mathcal{I}_1|} \sum_{\{i: i \in \mathcal{I}_1,\, W_i = 1\}} \left(\hat{\mu}_{(1)}^{\mathcal{I}_2}(X_i) - \mu_{(1)}(X_i) \right)^2} \\
&\qquad \times \sqrt{\frac{1}{|\mathcal{I}_1|} \sum_{\{i: i \in \mathcal{I}_1,\, W_i = 1\}} \left(\frac{1}{\hat{e}^{\mathcal{I}_2}(X_i)} - \frac{1}{e(X_i)} \right)^2} = o_P\left(\frac{1}{\sqrt{n}} \right)
\end{aligned}$$

by risk decay (3). (To establish this fact, also note that by consistency (2), the estimated propensities will all eventually also be uniformly bounded away from 0, $\eta/2 \leq \hat{e}^{\mathcal{I}_2}(X_i) \leq 1 - \eta/2$, and so the MSE for the inverse weights decays at the same rate as the MSE for the propensities themselves.)

The upshot is that by using cross-fitting, we can transform any $o_P(n^{-1/4})$-consistent machine learning method into an efficient ATE estimator. Also, the proof was remarkably short (at least compared to a typical proof in the semiparametric efficiency literature).

Condensed notation We will be encountering cross-fit estimators frequently in this class. From now on, we'll use the following notation: We define the data into K folds (above, $K = 2$), and compute estimators $\hat{\mu}_{(w)}^{(-k)}(x)$, etc., excluding the k-th fold. Then, writing $k(i)$ as the mapping that takes an observation and puts it into one of the k folds, we can write

$$\begin{aligned}
\hat{\tau}_{AIPW} = \frac{1}{n} \sum_{i=1}^{n} \Bigg(& \hat{\mu}_{(1)}^{(-k(i))}(X_i) - \hat{\mu}_{(0)}^{(-k(i))}(X_i) \\
& + W_i \frac{Y_i - \hat{\mu}_{(1)}^{(-k(i))}(X_i)}{\hat{e}^{(-k(i))}(X_i)} - (1 - W_i) \frac{Y_i - \hat{\mu}_{(0)}^{(-k(i))}(X_i)}{1 - \hat{e}^{(-k(i))}(X_i)} \Bigg),
\end{aligned} \tag{3.12}$$

which (almost) fits on one line.

Confidence intervals It is also important to be able to quantify uncertainty of treatment effect estimates. Cross-fitting also makes this easy. Recall from

last class that the empirical variance of the efficient score converges to the efficient variance V_*:

$$\frac{1}{n-1}\sum_{i=1}^{n}\left(\mu_{(1)}(X_i)-\mu_{(0)}(X_i) + W_i\frac{Y_i-\mu_{(1)}(X_i)}{e(X_i)}-(1-W_i)\frac{Y_i-\mu_{(0)}(X_i)}{1-e(X_i)}-\hat{\tau}^*\right)^2 \to_p V^*, \tag{3.13}$$

where $\hat{\tau}^*$ is as in (3.4). Our previous derivation then establishes that the same holds for cross-fitting: $\widehat{V}_{AIPW} \to_p V^*$, where

$$\widehat{V}_{AIPW} := \frac{1}{n-1}\sum_{i=1}^{n}\left(\hat{\mu}_{(1)}^{(-k(i))}(X_i)-\hat{\mu}_{(0)}^{(-k(i))}(X_i) + W_i\frac{Y_i-\hat{\mu}_{(1)}^{(-k(i))}(X_i)}{\hat{e}^{(-k(i))}(X_i)}-(1-W_i)\frac{Y_i-\hat{\mu}_{(0)}^{(-k(i))}(X_i)}{1-\hat{e}^{(-k(i))}(X_i)}-\hat{\tau}_{AIPW}\right)^2. \tag{3.14}$$

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

$$\tau \in \left(\hat{\tau}_{AIPW} \pm \frac{1}{\sqrt{n}}\Phi^{-1}(1-\frac{\alpha}{2})\sqrt{\widehat{V}_{AIPW}}\right),$$

where $\Phi(\cdot)$ is the standard Gaussian CDF, and these will achieve coverage with probability $1-\alpha$ in large samples. Similar argument can also be used to justify inference via resampling methods as in Efron [1982].

Closing thoughts People often ask whether using machine learning methods for causal inference necessarily means that our analysis becomes "uninterpretable." However, from a certain perspective, the results shown here may provide some counter evidence. We used "heavy" machine learning to obtain our estimates for $\hat{\mu}_{(w)}(x)$ and $\hat{e}(x)$—these methods were treated as pure black boxes, and we never looked inside—and yet the scientific questions we are trying to answer remain just as crisp as before (i.e., we want the ATE or the ATT). Perhaps our results even got *more* interpretable (or, at least, credible), because we did not need to rely on a parametric specification to build our estimators for τ.

Bibliographic references The literature on semiparametrically efficient treatment effect estimation via AIPW was pioneered by Robins, Rotnitzky, and

Zhao [1994], and developed in a sequence of papers including Robins and Rotnitzky [1995] and Scharfstein, Rotnitzky, and Robins [1999]. The effect of knowing the propensity score on the semiparametric efficiency bound for average treatment effect estimation is discussed in Hahn [1998], while the behavior of AIPW with high dimensional regression adjustments was first considered by Farrell [2015]. These results fit into a broader literature on semiparametrics, including Bickel, Klaassen, Ritov, and Wellner [1993] and Newey [1994].

The approach taken here, with a focus on generic machine learning estimators for nuisance components and cross-fitting, follows Chernozhukov et al. [2018a]. One major strength of this approach is in its generality and its ability to handle arbitrary nuisance estimators; however, the risk decay condition (3.8) is somewhat loose. There has been considerable recent interest in sharper analyses of AIPW that rely on specific choices of $\hat{\mu}_{(w)}(x)$ and $\hat{e}(x)$ to attain efficiency under the most general conditions possible, including work by Kennedy [2020] and Newey and Robins [2018].

Finally, one should note that AIPW is far from the only practical average treatment effect estimator that can attain semiparametric efficiency. One notable alternative to AIPW is targeted learning [van der Laan and Rubin, 2006], which can also be instantiated via machine learning based nuisance estimators and cross-fitting [van der Laan and Rose, 2011]. In the case of high-dimensional linear modeling, Belloni, Chernozhukov, and Hansen [2014] proposed a double-selection algorithm for choosing which variables to control for.

Lecture 4
Estimating Treatment Heterogeneity

Until now, we have focused on estimating the average treatment effect. In many application areas, however, there is interest in going beyond average effects, and to model treatment heterogeneity. For example, in personalized medicine, we may want to identify patients with more severe side effects than others; other application areas include public policy or online marketing. In this lecture, we'll discuss methods for treatment heterogeneity in observational studies that, analogously to AIPW, are to first order insensitive to errors in estimated nuisance components.

The conditional average treatment effect As always, we formalize our problem in terms of the potential outcomes framework. The analyst has access to n independent and identically distributed examples (X_i, Y_i, W_i), $i = 1, ..., n$, where $X_i \in \mathcal{X}$ denotes per-person features, $Y_i \in \mathbb{R}$ is the observed outcome, and $W_i \in \{0, 1\}$ is the treatment assignment. We posit the existence of potential outcomes $\{Y_i(0), Y_i(1)\}$ corresponding to the outcome we would have observed given the treatment assignment $W_i = 0$ or 1 respectively, such that $Y_i = Y_i(W_i)$.

In previous lectures, we focused on the average treatment effect $\tau = \mathbb{E}\left[Y_i(1) - Y_i(0)\right]$. Here, in contrast, we want to understand how treatment effects vary with the observed covariates X_i, and consider the conditional average treatment effect (CATE)

$$\tau(x) = \mathbb{E}\left[Y_i(1) - Y_i(0) \,\middle|\, X_i = x\right] \tag{4.1}$$

as our estimand. We emphasize that the CATE is not the same as the (in general unknowable) individual-i specific treatment effect $\Delta_i = Y_i(1) - Y_i(0)$; rather, it's still an average effect, but an average over a more targeted group of samples as characterized by their covariates X_i.

Regularization bias As discussed in the previous lecture that, whenever treatment assignment W_i is unconfounded, i.e., $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp W_i \,\big|\, X_i$, we can write

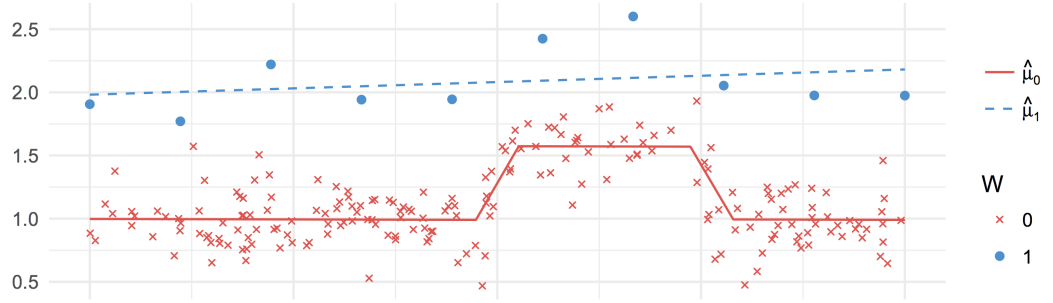
$$\tau(x) = \mu_{(1)}(x) - \mu_{(0)}(x), \qquad \mu_{(w)}(x) = \mathbb{E}\left[Y_i \,\middle|\, X_i = x,\, W_i = w\right]. \tag{4.2}$$

Now, since the $\mu_{(w)}(\cdot)$ are just two conditional response surfaces, one could imagine just fitting $\hat{\mu}_{(0)}(\cdot)$ and $\hat{\mu}_{(1)}(\cdot)$ by separate non-parametric regressions on the controls and treated units respectively, and then estimate the CATE as the difference between these two regression estimates,

$$\hat{\tau}_T(x) = \hat{\mu}_{(1)}(x) - \hat{\mu}_{(0)}(x). \tag{4.3}$$

This approach is simple and consistent (provided we use universally consistent estimators for $\mu_{(w)}(\cdot)$), but may not perform particularly well in finite samples.

A first concern is that, if there are many more control than treated units (or vice-versa) and we use generic non-parametric methods, then the two regression surfaces $\hat{\mu}_{(0)}(\cdot)$ and $\hat{\mu}_{(1)}(\cdot)$ may be differently regularized, thus creating artifacts in the learned CATE estimate $\hat{\tau}_T(x)$. The following figure, reproduced from Künzel, Sekhon, Bickel, and Yu [2019], illustrates this point. Both $\mu_{(0)}(x)$ and $\mu_{(1)}(x)$ vary with x but the CATE function is constant. There are many controls so $\hat{\mu}_{(0)}(\cdot)$ is well estimated, but there are very few treated treated units and so $\hat{\mu}_{(1)}(\cdot)$ is heavily regularized and approximated as a linear function. Both estimates $\hat{\mu}_{(0)}(\cdot)$ and $\hat{\mu}_{(1)}(\cdot)$ are reasonable on their own; however, once we take their difference as in (4.3), we find strong heterogeneity is $\tau(x)$ where there is none (which is effectively the worst thing a statistical method can do).



A second, more subtle concern is that (4.3) does not explicitly account for variation in the propensity score. If $e(x)$ varies considerably, then our estimates of $\hat{\mu}_{(0)}(\cdot)$ will be driven by data in areas with many control units (i.e., with $e(x)$ closer to 0), and those of $\hat{\mu}_{(1)}(\cdot)$ by regions with more treated units (i.e., with $e(x)$ closer to 1). And if there is covariate shift between the data used to learn $\hat{\mu}_{(0)}(\cdot)$ and $\hat{\mu}_{(1)}(\cdot)$, this may create biases for their difference $\hat{\tau}_T(x)$.

Semiparametric modeling In order to develop a more formal understanding of heterogeneous treatment effect estimation, it is helpful to consider the case where we have a model for $\tau(x)$,

$$Y_i(w) = f(X_i) + w\,\tau(X_i) + \varepsilon_i(w), \quad \mathbb{P}\left[W_i = 1 \,\middle|\, X_i\right] = e(x), \tag{4.4}$$

where $\tau(x) = \psi(x) \cdot \beta$ for some pre-determined set of basis functions $\psi : \mathcal{X} \to \mathbb{R}^k$. In other words, we allow for non-parametric relationships between X_i, Y_i, and W_i; however, the treatment effect function itself is parametrized by $\beta \in \mathbb{R}^k$.

This class of problems was studied by Robinson [1988] who showed that, under unconfoundedness, we can re-write (4.4) as

$$\begin{aligned} &Y_i - m(X_i) = (W_i - e(X_i))\,\psi(X_i) \cdot \beta + \varepsilon_i, \text{ where} \\ &m(x) = \mathbb{E}\left[Y_i \,\middle|\, X_i = x\right] = f(X_i) + e(X_i)\tau(X_i) \end{aligned} \tag{4.5}$$

denotes the conditional expectation of the observed Y_i, marginalizing over W_i and $\varepsilon_i = \varepsilon_i(w)$.

This suggests the following "oracle" algorithm for estimating β: First define $\widetilde{Y}_i^* = Y_i - m(X_i)$ and $\widetilde{Z}_i^* = (W_i - e(X_i))\psi(X_i)$, and then estimate $\hat{\zeta}_R^*$ by running residual-on-residual OLS regression $\widetilde{Y}_i^* \sim \widetilde{Z}_i^*$. One can show that this oracle procedure is $\sqrt{n}$-consistent and asymptotically normal,[1]

$$\sqrt{n}\left(\hat{\zeta}^* - \beta\right) \Rightarrow \mathcal{N}\left(0,\, V_R\right), \quad V_R = \text{Var}\left[\widetilde{Z}_i^*\right]^{-1} \text{Var}\left[\widetilde{Z}_i^* \widetilde{Y}_i^*\right] \text{Var}\left[\widetilde{Z}_i^*\right]^{-1}. \tag{4.6}$$

Moreover, under homoskedasticity, i.e., in the case where $\text{Var}\left[\varepsilon_i \,\middle|\, X_i,\, W_i\right] = \sigma^2$ is constant, V_R is the semiparametrically efficient variance for estimating β (under heteroskedasticity, result (4.6) still holds, but V_R is no longer the semiparametrically efficient variance).

We of course can't use this oracle estimator in practice since we don't know $m(x)$ and $e(x)$. However, we can again use cross fitting to emulate the oracle:

1. Run non-parametric regressions $Y \sim X$ and $W \sim X$ using a method of our choice to get $\hat{m}(x)$ and $\hat{e}(x)$ respectively.

2. Define transformed features $\widetilde{Y}_i = Y_i - \hat{m}^{(-k(i))}(X_i)$ and $\widetilde{Z}_i = (W_i - \hat{e}^{(-k(i))}(X_i))\psi(X_i)$, using cross-fitting for $\hat{m}(x)$ and $\hat{e}(x)$ as usual.

3. Estimate $\hat{\zeta}_R$ by running the OLS regression $\widetilde{Y}_i \sim \widetilde{Z}_i$.

[1] For a recent review of OLS asymptotics without linear modeling assumptions, see Buja et al. [2019].

Using a similar argument as discussed in class last time, we can verify that if all non-parametric regressions satisfy

$$\mathbb{E}\left[(\hat{m}(X) - m(X))^2\right]^{\frac{1}{2}},\ \mathbb{E}\left[(\hat{e}(X) - e(X))^2\right]^{\frac{1}{2}} = o_P\left(\frac{1}{n^{1/4}}\right), \tag{4.7}$$

then cross-fitting emulates the oracle,

$$\sqrt{n}(\hat{\zeta}_R^* - \hat{\zeta}_R) \to_p 0 \tag{4.8}$$

and so $\hat{\zeta}_R$ has the same distribution as in (4.6); see Chernozhukov et al. [2018a] for details.

A loss function for treatment heterogeneity The estimator of Robinson for the partially linear model (4.4) provides helpful guidance on how to derive robust estimates of the CATE in observational studies if we are willing to use a linear specification $\tau(x) = \psi(x) \cdot \beta$. In many modern setting with complex covariates, however, we may not want to commit to a linear form for $\tau(x)$ a-priori, and would prefer to use a machine learning method that can adaptively discover a good representation for the CATE.

To this end, it is helpful to re-write Robinson's estimator as a loss minimizer. Writing conditional response surfaces as $\mu_{(w)}(x) = \mathbb{E}\left[Y(w) \,\middle|\, X = x\right]$ for $w \in \{0, 1\}$ we observe that, under unconfoundedness,

$$\mathbb{E}\left[\varepsilon_i(W_i) \mid X_i,\, W_i\right] = 0, \text{ where } \varepsilon_i(w) := Y_i(w) - \left(\mu_{(0)}(X_i) + w\tau(X_i)\right). \tag{4.9}$$

We can then follow Robinson's approach, and re-write

$$Y_i - m(X_i) = (W_i - e(X_i))\,\tau\,(X_i) + \varepsilon_i, \tag{4.10}$$

where $m(x) = \mathbb{E}\left[Y \,\middle|\, X = x\right] = \mu_{(0)}(X_i) + e(X_i)\tau(X_i)$ and $\varepsilon_i := \varepsilon_i(W_i)$ (note that this decomposition holds for any outcome distribution, including for binary outcomes).

Furthermore, (4.10) can equivalently be expressed as

$$\tau(\cdot) = \text{argmin}_{\tau'}\left\{\mathbb{E}\left[\left(\,(Y_i - m(X_i)) - (W_i - e(X_i))\,\tau'(X_i)\right)^2\right]\right\}, \tag{4.11}$$

and so an oracle who knew both the functions $m(x)$ and $e(x)$ a priori could estimate the heterogeneous treatment effect function $\tau(\cdot)$ by empirical loss

minimization,

$$\hat{\tau}_R^*(\cdot) = \text{argmin}_{\tau'} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left((Y_i - m(X_i)) - (W_i - e(X_i))\, \tau'(X_i) \right)^2 + \Lambda_n \left(\tau'(\cdot) \right) \right\}, \tag{4.12}$$

where the term $\Lambda_n(\tau(\cdot))$ is interpreted as a regularizer on the complexity of the $\tau(\cdot)$ function. In practice, this regularization could be explicit as in penalized regression such as the lasso or kernel regression, or implicit, e.g., as provided by a carefully designed deep neural network.

The difficulty, as always, is that in practice we never know the weighted main effect function $m(x)$ and usually don't know the treatment propensities $e(x)$ either (unless we're in an RCT), and so the estimator (4.12) is not feasible. Thus, it's natural to consider a plug-in alternative via cross-fitting

$$\begin{aligned} \hat{\tau}_R(\cdot) &= \text{argmin}_{\tau} \left\{ \widehat{L}_n \left(\tau(\cdot) \right) + \Lambda_n \left(\tau(\cdot) \right) \right\}, \\ \widehat{L}_n \left(\tau(\cdot) \right) &= \frac{1}{n} \sum_{i=1}^{n} \left(\left(Y_i - \hat{m}^{(-k(i))}(X_i) \right) - \left(W_i - \hat{e}^{(-k(i))}(X_i) \right) \tau(X_i) \right)^2. \end{aligned} \tag{4.13}$$

There are many ways to act on the estimation strategy. For example, using the lasso, (4.13) becomes $\hat{\tau}(x) = x \cdot \hat{\beta}$ with[2]

$$\hat{\beta} = \text{argmin}_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\left(Y_i - \hat{m}^{(-k(i))}(X_i) \right) - \left(W_i - \hat{e}^{(-k(i))}(X_i) \right) X_i \beta \right)^2 + \lambda \left\| \beta \right\|_1 \right\}; \tag{4.14}$$

or, one could directly use $\widehat{L}_n(\cdot)$ as a loss function for boosting or deep learning. Nie and Wager [2017] establish conditions where (4.13) has a quasi-oracle property analogous to the one discussed above, and $\hat{\tau}_R$ can emulate the best performance guarantees available for the oracle $\hat{\tau}_R^*$; following this paper, we refer to $\widehat{L}_n$ as the R-loss.

[2]In some cases, it appears that adding a main effect term to the lasso problem (4.14), i.e., running a penalized regression with $\left(Y_i - \hat{m}^{(-k(i))}(X_i) \right) \sim X_i \zeta + \left(W_i - \hat{e}^{(-k(i))}(X_i) \right) X_i \beta$, may improve empirical performance slightly [Nie and Wager, 2017].

Validating treatment heterogeneity When working with flexible approaches to estimating treatment heterogeneity, it's important to be able to rigorously validate and choose between candidate estimators. How best to validate treatment effect estimators is still a somewhat open topic; however, several possible solutions have been proposed in the literature.

A first, simple approach that builds directly on (4.13) is to cross-validate on the R-loss, i.e., prefer the estimator with the smallest out-of-fold R-loss.[3] Furthermore, working in a loss-minimization framework opens the door to a broader machine learning approach. In addition to using the R-loss $\widehat{L}_n(\cdot)$ for choosing between competing estimators via cross-validation, one could also use it for, e.g., model aggregation via stacking or pruning a complex model of treatment heterogeneity [van der Laan, Polley, and Hubbard, 2007].

Another, more indirect approach is to use a conditional average treatment effect estimator $\hat{\tau}(x)$ to guide subgroup analysis. For example, one could stratify a test set according to estimates of $\hat{\tau}(x)$, and then estimate the average treatment effect separately for each stratum using doubly robust methods as discussed in Lecture 3; then, a good treatment effect estimator is one that can reproducibly find subgroups with different average treatment effects (of course, for this to be valid, the data used for estimating the ATEs over subgroups cannot be the same as the data used to learn $\hat{\tau}(x)$).

Finally, if one is simply interested in validation, one could—again on a hold-out set—try fitting a partially linear model as in (4.10), but with treatment effect function parametrized in terms of the estimated CATE function, i.e., $\tau(x) \sim \alpha + \beta\hat{\tau}(x)$. If we run Robinson's method with this parametrization and find $\hat{\beta} \approx 1$, this may be taken as evidence that the treatment heterogeneity estimate is well calibrated; meanwhile, if $\hat{\beta}$ is significantly greater than 0, this may be taken as evidence that our estimated CATE function $\hat{\tau}(x)$ is not pure noise. For further examples and discussion, see Athey and Wager [2019] and Chernozhukov, Demirer, Duflo, and Fernández-Val [2017].

Closing thoughts At first glance, the problem of estimating treatment heterogeneity may seem like just another non-parametric regression problem: Just learn $\hat{\mu}_{(w)}(x)$ as usual, and then estimate the CATE via (4.3). However, regularization bias (meaning biases that arise from poorly targeted objectives for the

[3]One practical issue that may arise when using the R-loss for model choice is that the numerical difference between the cross-validated losses may be very small relative to the sampling error of the R-loss itself. Somewhat surprisingly, this may not always be a problem due to a general phenomenon with cross-validation, whereby the leading noise term of the cross-validated error cancels out when we compare two models; see Wager [2020a] for a discussion.

treatment and control models) can be a real problem if not addressed up front. These difficulties are particularly accute in the case of causal effect estimation, because we are often interested in estimating potentially weak treatment effects $\tau(x)$ in the presence of much stronger baseline effects $\mu_{(0)}(x)$ (e.g., in a medical application, one might expect that the causal effect of any intervention on survival is much smaller than baseline variation in survival probabilities across patients).

An early line of work on methods for treatment heterogeneity sought to address regularization bias by directly modifying popular statistical learning tools such as the lasso or regression trees to focus on accurate estimation of the CATE in randomized trials [Athey and Imbens, 2016, Imai and Ratkovic, 2013, Tian, Alizadeh, Gentles, and Tibshirani, 2014]. Here, in contrast, we saw how an extension of the partial linear model estimator of Robinson can be used to create a loss function that directly targets the CATE function; and then we can get good estimates of $\tau(\cdot)$ by simply minimizing this loss.

The key fact that enabled the whole approach discussed here is the "quasi oracle" property (4.8) for Robinson's method, according to the feasible version of Robinson's estimator with estimated nuisance components $\hat{e}(x)$ and $\hat{m}(x)$ is to first order just as good as the oracle with known nuisance components—provided the condition (4.7) holds. This result is closely related to the robustness property of AIPW discussed in the last lecture, where again errors in the nuisance components didn't matter to first order. Such estimators, which Chernozhukov et al. [2018a] refer to as Neyman-orthogonal, play a key role in non-parametric causal inference (and semiparametric statistics more broadly).

Bibliographic notes Today, we discussed an approach to heterogeneous treatment effect estimation in observational studies that builds on the estimator of Robinson [1988] for partially linear modeling. In further results in this line of work, Nie and Wager [2017] present excess error bounds for heterogeneous treatment effect estimation via non-parametric kernel regression with the R-learner, and Zhao, Small, and Ertefaie [2017] discuss post-selection inference for treatment heterogeneity using what we've here called the R-lasso. A random forest based variant of the R-learner is implemented in the `causal_forest` function in the `R`-package `grf` [Athey, Tibshirani, and Wager, 2019].

Other recently proposed methods for heterogeneous treatment effect estimation in observational studies include Hahn, Murray, and Carvalho [2020] and Künzel, Sekhon, Bickel, and Yu [2019], who propose different approaches using the propensity score for this problem (although these methods are not orthogonal to errors in nuisance components in the sense of (4.8)). Finally, Ding, Feller, and Miratrix [2019] discuss estimation of treatment heterogeneity in a

randomized trial under strict randomization inference (i.e., without assuming a sampling distribution for the potential outcomes).

As an aside, we note that Robinson's estimator for the partial linear model can also be of interest when the partially linear model is misspecified. In the simplest case where treatment effects are constant, i.e., $Y_i(w) = f(X_i) + \tau w + \varepsilon_i$ and $\mathbb{E}[\varepsilon_i \,\big|\, X_i, W_i]$, Robinson's method provides a simple and consistent estimate of the treatment effect parameter τ provided nuisance components converge fast enough:

$$\hat{\tau}_R = \frac{\sum_{i=1}^n \left(Y_i - \hat{m}^{(-k(i))}(X_i)\right)\left(W_i - \hat{e}^{(-k(i))}(X_i)\right)}{\sum_{i=1}^n \left(W_i - \hat{e}^{(-k(i))}(X_i)\right)^2}. \tag{4.15}$$

However, even when the conditional average treatment effect function $\tau(x) = \mathbb{E}[Y_i(1) - Y_i(0) \,\big|\, X_i = x]$ is not constant, one can verify that $\hat{\tau}_R$ converges to a weighted average of $\tau(x)$ with non-negative weights, and that $\hat{\tau}_R$ is substantially more robust to local failures of overlap than efficient estimators of the average treatment effect [Crump, Hotz, Imbens, and Mitnik, 2009, Li, Morgan, and Zaslavsky, 2018]. Thus, in cases where we believe heterogeneity in $\tau(x)$ to be low and we have difficulties with overlap, using (4.15) as an alternative to an average treatment effect estimator may be a practical choice.

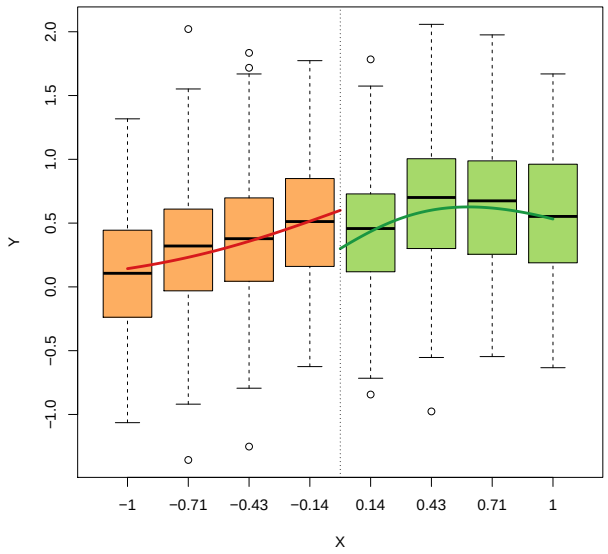
Lecture 5
Regression Discontinuity Designs

The cleanest and most straight-forward approach to treatment effect estimation is via the randomized controlled trial and its immediate generalizations. However, in applied work, there are several other quasi-experimental designs that have repeatedly proven themselves in practice. One simple yet versatile approach of this type is the regression discontinuity design, which relies on discontinuous treatment assignment mechanisms to identify causal effects. Today, we'll formalize identification arguments for regression discontinuity designs, and discuss best practices for estimation.

Setting and motivation We are interested in the effect of a binary treatment W_i on a real-valued outcome Y_i, and posit potential outcomes $\{Y_i(0), Y_i(1)\}$ such that $Y_i = Y_i(W_i)$. However, unlike in a randomized trial, we do not take the treatment assignment W_i to be random. Instead, we assume there is a running variable $Z_i \in \mathbb{R}$ and a cutoff c, such that $W_i = 1\left(\{Z_i \geq c\}\right)$. This setting could arise, e.g., in education, where Z_i is a standardized test score and students with $Z_i \geq c$ are eligible to enroll in an honors program, or in medicine, where Z_i is a severity score, and patients are prescribed an intervention once $Z_i \geq c$.

Qualitatively, the main idea of a regression discontinuity is that although treatment assignment W_i is not randomized, it's almost as good as random when Z_i is in the vicinity of the cutoff c. People with Z_i close to c ought to all be similar to each other on average, but only those with $Z_i \geq c$ get treated, and so we can estimate a treatment effect by comparing people with Z_i right above versus right below 0.

Identification via continuity The most prevalent way to formalize the qualitative argument made above is by invoking continuity. Let $\mu_{(w)}(z) = \mathbb{E}\left[Y_i(w) \,\middle|\, Z_i\right]$. Then, if $\mu_{(0)}(z)$ and $\mu_{(1)}(z)$ are both continuous, we can identify



the conditional average treatment effect at $z = c$, i.e., $\tau_c = \mu_{(1)}(c) - \mu_{(0)}(c)$, via

$$\tau_c = \lim_{z \downarrow c} \mathbb{E}\left[Y_i \,\middle|\, Z_i = z\right] - \lim_{z \uparrow c} \mathbb{E}\left[Y_i \,\middle|\, Z_i = z\right], \tag{5.1}$$

provided that the running variable Z_i has support around the cutoff c. In other words, we identify τ_c as the difference between the endpoints of two different regression curves; the above figure provides an illustration.

Why our previous results don't apply to RDD Before discussing methods for estimation in regression discontinuity designs, it's helpful to consider why our previously considered approaches (such as IPW) don't apply. As emphasized by Rubin [2008], the two assumptions that are invariably needed in studying quasi-experiments (and were used throughout our discussion so far) are

$$\{Y_i(0),\, Y_i(1)\} \perp\!\!\!\perp W_i \,\big|\, Z_i, \qquad \text{(unconfoundedness)} \tag{5.2}$$

$$\eta \leq \mathbb{P}\left[W_i = 1 \,\middle|\, Z_i\right] \leq 1 - \eta, \qquad \text{(overlap)} \tag{5.3}$$

for some $\eta > 0$. Taken together, unconfoundedness and overlap mean that we can view our dataset as formed by pooling many small randomized trials indexed by different values of Z_i; then, unconfoundedness means that treatment

assignment is exogenous given Z_i, while overlap means that randomization in fact occurred (e.g., one can't learn anything from a randomized trial where everyone is assigned to control).

In a regression discontinuity design, we have $W_i = 1\left(\{Z_i \geq c\}\right)$, and so unconfoundedness holds trivially (because W_i is a deterministic function of Z_i). However, overlap clearly doesn't hold: $\mathbb{P}\left[W_i = 1 \,\middle|\, Z_i = z\right]$ is always either 0 or 1. Thus, methods like IPW that involve division by $\mathbb{P}\left[W_i = 1 \,\middle|\, Z_i\right]$, etc., are not applicable. Instead, we'll need to compare units with Z_i straddling the cutoff c that are similar to each other—but do not have contiguous distributions.

On a statistical level, the main consequence of this failure of overlap is that $\sqrt{n}$-consistent estimation of τ_c is in general not possible. Instead, the minimax error for estimating τ_c will decay at sub-parametric rates, and the specific rate will depend on how smooth the conditional response functions $\mu_{(w)}(z)$ are. For example, if the $\mu_{(w)}(z)$ have a uniformly bounded second derivative in the vicinity of the cutoff c then, as we'll see below, we can achieve $n^{-2/5}$ rates of convergence.

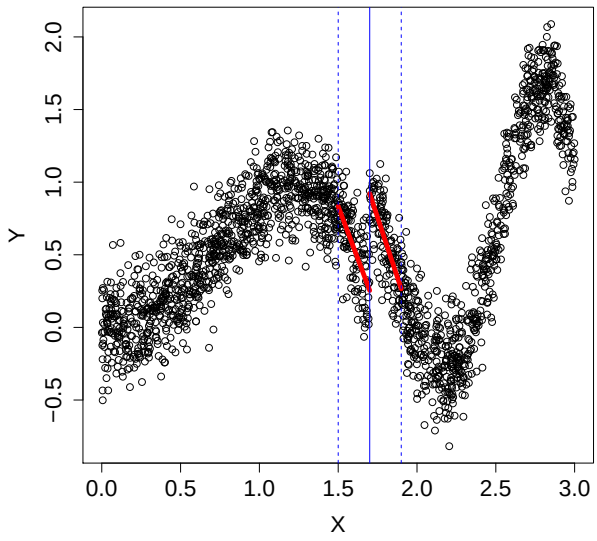
Estimation via local linear regression A simple and robust approach to estimation based on (5.1) is to use local linear regression, as illustrated in the figure below. We pick a small bandwidth $h_n \to 0$ and a symmetric weighting function $K(\cdot)$, and then fit $\mu_{(w)}(z)$ via weighted linear regression on each side of the boundary,

$$\begin{aligned}\hat{\tau}_c = \text{argmin} \Bigg\{ \sum_{i=1}^{n} K\left(\frac{|Z_i - c|}{h_n}\right) \\ \times \left(Y_i - a - \tau W_i - \beta_{(0)}\left(Z_i - c\right)_- - \beta_{(1)}\left(Z_i - c\right)_+\right)^2 \Bigg\},\end{aligned} \tag{5.4}$$

where the overall intercept a and slope parameters $\beta_{(w)}$ are nuisance parameters. Popular choices for the weighting function $K(x)$ include the window function $K(x) = 1\left(\{|x| \leq 1\}\right)$, or the triangular kernel $K(x) = \left(1 - |x|\right)_+$.

Consistency, asymptotics and rates of convergence It is not hard to see that, under continuity assumptions as in (5.1), the local linear regression estimator (5.4) must be consistent for reasonable choices of the bandwidth sequence h_n. However, in order to move beyond such a high-level statement and get any quantitative guarantees, we need to be more specific about the continuity assumptions made on $\mu_{(0)}(z)$ and $\mu_{(1)}(z)$.

There are many ways of quantifying smoothness, but one of the most widely used assumptions in practice—and the one we'll focus on today—is that the



$\mu_{(w)}(z)$ are twice differentiable with a uniformly bounded second derivative

$$\left|\frac{d^2}{dz^2}\mu_{(w)}(z)\right| \leq B \text{ for all } z \in \mathbb{R} \text{ and } w \in \{0,1\}\,. \tag{5.5}$$

One motivation for the assumption (5.5) is that it justifies local linear regression as in (5.4): If we had less smoothness (e.g., $\mu_{(w)}(z)$ is just taken to be Lipschitz) then there would be no point doing local linear regression as opposed to local averaging, whereas if we had more smoothness (e.g., bounds on the k-th order derivative of $\mu_{(w)}(z)$ for $k \geq 3$) then we could improve rates of convergence via local regression with higher-order polynomials.

Given this assumption, we can directly bound the error rate of (5.4). First, by taking a Taylor expansion around c, we can write

$$\mu_{(w)}(z) = a_{(w)} + \beta_{(w)}(z-c) + \frac{1}{2}\rho_{(w)}(z-c), \qquad \left|\rho_{(w)}(x)\right| \leq Bx^2, \tag{5.6}$$

while noting that $\tau_c = a_{(1)} - a_{(0)}$. Moreover, by inspection of the problem (5.4), we see that it factors into two separate regression problems on the treated and control samples, namely

$$\hat{a}_{(1)},\ \hat{\beta}_{(1)} = \text{argmin}_{a,\beta}\left\{\sum_{Z_i \geq c} K\left(\frac{|Z_i - c|}{h_n}\right)\left(Y_i - a - \beta\left(Z_i - c\right)\right)^2\right\}, \tag{5.7}$$

for the treated units and an analogous problem for the controls, such that $\hat{\tau} = \hat{a}_{(1)} - \hat{a}_{(0)}$.

Now, for simplicity, focus on local linear regression with the basic window kernel $K(x) = 1\left(\{|x| \leq 1\}\right)$. The linear regression problem (5.7) can then be solved in closed form, and we get

$$\hat{a}_{(1)} = \sum_{c \leq Z_i \leq c+h_n} \gamma_i Y_i, \quad \gamma_i = \frac{\widehat{\mathbb{E}}_{(1)}\left[(Z_i - c)^2\right] - \widehat{\mathbb{E}}_{(1)}\left[Z_i - c\right] \cdot (Z_i - c)}{\widehat{\mathbb{E}}_{(1)}\left[(Z_i - c)^2\right] - \widehat{\mathbb{E}}_{(1)}\left[Z_i - c\right]^2}, \tag{5.8}$$

where $\widehat{\mathbb{E}}_{(1)}\left[Z_i - c\right] = \sum_{c \leq Z_i \leq c+h_n} (Z_i - c) / \left|\{i : c \leq Z_i \leq c + h_n\}\right|$, etc., denote sample averages over the regression window. Now, by direct calculation we see that $\sum_{c \leq Z_i \leq c+h_n} \gamma_i = 1$ and $\sum_{c \leq Z_i \leq c+h_n} \gamma_i (Z_i - c) = 0$ and so, thanks to (5.6), we see that

$$\hat{a}_{(1)} = a_{(1)} + \underbrace{\sum_{c \leq Z_i \leq c+h_n} \gamma_i \, \rho_{(1)}(Z_i - c)}_{\text{curvature bias}} + \underbrace{\sum_{c \leq Z_i \leq c+h_n} \gamma_i \, \left(Y_i - \mu_{(1)}(Z_i)\right)}_{\text{sampling noise}}, \tag{5.9}$$

and a similar expansion holds for $\hat{a}_{(0)}$. Thus, recalling that our estimator is $\hat{\tau} = \hat{a}_{(1)} - \hat{a}_{(0)}$ and out target estimand is $\tau_c = a_{(1)} - a_{(0)}$, we see that it suffices to bound the error terms in (5.9).

Given our bias on the curvature, we immediately see that the "curvature bias" term is bounded by Bh_n^2. Meanwhile, the sampling noise term is mean-zero and, provided that $\mathrm{Var}\left[Y_i \,\middle|\, Z_i\right] \leq \sigma^2$, has variance bounded on the order of $\sigma^2 \sum_{c \leq Z_i \leq c+h_n} \gamma_i^2$. Finally, assuming that Z_i has a continuous non-zero density function $f(z)$ in a neighborhood of z, one can check that

$$\sigma^2 \sum_{c \leq Z_i \leq c+h_n} \gamma_i^2 \approx \frac{4\sigma^2}{\left|\{i : c \leq Z_i \leq c + h_n\}\right|} \approx \frac{4\sigma^2}{f(c)} \frac{1}{n h_n}. \tag{5.10}$$

In other words, the squared bias of $\hat{\tau}$ scales as h_n^4, while its variance scales as $1/(h_n n)$. The bias-variance trade-off is minimized by tuning h_n, and we find that

$$\hat{\tau}_c = \tau_c + \mathcal{O}_P\left(n^{-2/5}\right), \quad \text{with} \quad h_n \sim n^{-1/5}. \tag{5.11}$$

In other words, we have established that if the potential outcome functions have bounded curvature as in (5.5) and Z_i has a continuous non-zero density around c (meaning that there will asymptotically be datapoints with Z_i arbitrarily close to c), then local linear regression can estimate τ_c at an $n^{-2/5}$ rate.

Finally, note that this $n^{-2/5}$ rate is a consequence of working with bounds on the 2nd derivative of $\mu_{(w)}(z)$. In general, if we assume that $\mu_{(w)}(z)$ has

a bounded k-th order derivative, then we can achieve an $n^{-k/(2k+1)}$ rate of convergence for τ_c by using local polynomial regression of order $(k-1)$ with a bandwidth scaling as $h_n \sim n^{-1/(2k+1)}$. Local linear regression never achieves a parametric rate of convergence, but can get more or less close depending on how smooth $\mu_{(w)}(z)$ is.

Identification via noisy running variables So far, we have focused on identification in regression discontinuity designs via continuity of $\mu_{(w)}(x)$; specifically, we assumed that the second derivative of $\mu_{(w)}(x)$ is bounded as in (5.5). However, despite its simplicity and interpretability, this continuity-based approach to regression discontinuity inference does not satisfy the criteria for rigorous design-based causal inference as outlined by Rubin [2008]. According to the design-based paradigm, even in observational studies, a treatment effect estimator should be justifiable based on randomness in the treatment assignment mechanism alone. In contrast, the formal guarantees provided by the continuity-based regression discontinuity analysis take smoothness of $\mu_{(w)}(z)$ as a primitive.

An alternative justification for identification in regression discontinuity designs starts with a form of implicit randomization in the running variable: There are many factors outside of the control of decision-makers that determine the running variable Z_i such that if some unit barely clears the eligibility cutoff for the intervention then the same unit could also plausibly have failed to clear the cutoff with a different realization of these chance factors [Lee and Lemieux, 2010]. For example, in an educational setting where a test is used to determine eligibility to an honors program, there may be a group of marginal students who might barely pass or fail pass a test due to unpredictable variation in their test score, thus resulting in an effectively exogenous treatment assignment rule.

And, if the running variable is in fact noisy, we can build an identification argument on top of it.[1] More formally, consider a setting where the following two conditions hold:

- The running variable is noisy, such that there is a latent variable U_i with distribution G such that $Z_i \,\big|\, U_i \sim \mathcal{N}\left(0,\, \nu^2\right)$ for some $\nu > 0$.

[1]Running variables are plausibly noisy in many, but not all, applications of RDDs. For example, in some cases one might consider an RDD where different counties enact different policies and so there is a sharp cutoff in legislation at the county border. Here, a continuity-based argument may be applicable, but claiming that a household's position in space is noisy seems questionable.

- The noise in Z_i is unconfounded or exogenous, i.e., $[\{Y_i(0),\, Y_i(1)\} \perp\!\!\!\perp Z_i] \,\big|\, U_i$.

Once we invoke this latent structure, we recover an average treatment effect estimation problem that's reminiscent from the one studied last week: We have $Y_i = Y_i(W_i)$, with $\{Y_i(0),\, Y_i(1)\} \perp\!\!\!\perp W_i \,\big|\, U_i$ and

$$\begin{aligned} e(u) &:= \mathbb{P}\left[W_i = 1 \,\middle|\, U_i\right] = \mathbb{P}\left[Z_i \geq c \,\middle|\, U_i\right] = 1 - \Phi\left(\frac{c - U_i}{\nu}\right), \\ \alpha_{(w)}(u) &:= \mathbb{E}\left[Y_i(w) \,\middle|\, U_i = u\right], \quad \tau(u) = \mathbb{E}\left[Y_i(1) - Y_i(0) \,\middle|\, U_i = u\right], \end{aligned} \tag{5.12}$$

where $\Phi(\cdot)$ is the standard Gaussian cumulative distribution function.

The remaining difficulty, of course, is that the latent variable U_i is not observed and so we cannot control for it as in, e.g., AIPW. However, as discussed further in Eckles, Ignatiadis, Wager, and Wu [2020], one can address this issue using a deconvolution-type estimator. Specifically, if one sets

$$\hat{\tau}_\gamma = \frac{1}{n} \sum_{\{i: Z_i \geq c\}} \gamma_+(Z_i) Y_i - \frac{1}{n} \sum_{\{i: Z_i < c\}} \gamma_-(Z_i) Y_i \tag{5.13}$$

with weighting functions satisfying

$$\mathbb{E}\left[1\left(\{Z_i \geq c\}\right) \gamma_+(Z_i)\right] = \mathbb{E}\left[1\left(\{Z_i < c\}\right) \gamma_-(Z_i)\right] = 1, \tag{5.14}$$

then one can verify that

$$\begin{aligned} \mathbb{E}\left[\hat{\tau}_\gamma\right] &= \underbrace{\int h_+(u) \tau(u)\, dG(u)}_{\text{weighted treatment effect}} + \underbrace{\int \left(h_+(u) - h_-(u)\right) \alpha_{(0)}(u)\, dG(u)}_{\text{confounding bias}}, \\ h_+(u) &= \int_c^\infty \gamma_+(z) \varphi_\nu(z - u)\, dz, \quad h_-(u) = \int_{-\infty}^c \gamma_-(z) \varphi_\nu(z - u)\, dz, \end{aligned} \tag{5.15}$$

and this path can be further pursued to devise estimators for various weighted averages of $\tau(u)$ as defined in (5.12).

Bibliographic notes The idea of using regression discontinuity designs for treatment effect estimation goes back to Thistlethwaite and Campbell [1960]; however, most formal work in this area is more recent. The framework of identification in regression discontinuity designs via continuity arguments and local linear regression is laid out by Hahn, Todd, and van der Klaauw [2001]. Other references on regression discontinuity analysis via local linear regression include Cheng, Fan, and Marron [1997] who discuss optimal choices for

the kernel weighting function, Imbens and Kalyanaraman [2012] who discuss bandwidth choice, and Calonico, Cattaneo, Farrell, and Titiunik [2019] who discuss the role of covariate adjustments. Imbens and Lemieux [2008] provide an overview of local linear regression methods in this setting, and discuss alternative specifications such as the "fuzzy" regression discontinuities where W_i is random but $\mathbb{P}\left[W_i = 1 \,\middle|\, Z_i = z\right]$ has a jump at the cutoff c.

One topic we did not discuss today is the construction of confidence intervals via local linear regression. The reason this is a somewhat delicate issue is that, when tuned for optimal mean-squared error, the bias and sampling error of the local linear regression estimator are of the same order, and so basic delta-method or bootstrap based inference fails (because it doesn't capture bias). Several authors have considered solutions to the problem that rely on asymptotics. In particular, Calonico, Cattaneo, and Titiunik [2014] and Calonico, Cattaneo, and Farrell [2018] bias-correct local linear regression to obtain valid confidence intervals, while Armstrong and Kolesár [2020] show that uncorrected local linear regression point estimates can also be used for valid inference provided we inflate the length of the confidence intervals by a pre-determined amount. We will revisit the problem of inference for regression discontinuity designs in the next lecture, with a focus on methods that allow for finite sample justification.

When discussing alternative identification via noisy running variables, we found it helpful to consider conditioning on an unobserved latent variable U_i to study the behavior of our estimator. This idea, sometimes called principle stratification, plays an important role in many key results about non-parametric causal inference in observational studies [Frangakis and Rubin, 2002, Heckman and Vytlacil, 2005, Imbens and Angrist, 1994], and we will encounter it again when working with instrumental variables.

Lecture 6
Finite Sample Inference in RDDs

In the previous lecture, we introduced regression discontinuity designs as a strategy for identifying causal effects. To recap, we asked about the effect of a binary treatment W_i on a real-valued outcome Y_i in a setting where treatment assignment is a deterministic function of a continuous running variable $Z_i \in \mathbb{R}$, i.e., there is a cutoff c, such that $W_i = 1\left(\{Z_i \geq c\}\right)$. Assuming potential outcomes $\{Y_i(0), Y_i(1)\}$ such that $Y_i = Y_i(W_i)$ we found that, under continuity assumptions, $\tau_c = \mathbb{E}\left[Y_i(1) - Y_i(0) \,\middle|\, Z_i = c\right]$ is identified via

$$\tau_c = \lim_{z \downarrow c} \mathbb{E}\left[Y_i \,\middle|\, Z_i = z\right] - \lim_{z \uparrow c} \mathbb{E}\left[Y_i \,\middle|\, Z_i = z\right]. \tag{6.1}$$

Furthermore, we showed that a simple estimator based on local linear regression can achieve an $n^{-2/5}$ rate of convergence if the conditional response functions of the treated and control potential outcomes have bounded second derivatives.

Now, while this result is very helpful from a conceptual point of view, it is not always clear how to use it in practice. In particular:

- The asymptotic argument underlying (6.1) relies on observing data Z_i arbitrarily close to the cutoff c. In practice, however, we often have to work with discrete running variables (e.g., Z_i is a test score that takes integers value between 0 and 100), and so these asymptotics do not apply.

- When Z_i has a continuous distribution and we run local linear regression with an optimal bandwidth, both the bias and standard error of $\hat{\tau}_c$ are of the same order of magnitude. Thus, any approach to inference that does not account for bias won't achieve coverage.

- In many applications, we need to work with more complicated cutoff functions (e.g., a student needs to pass 2 out of 3 tests to be eligible for a program). How does (6.1) generalize to this setting?

Our goal today is to re-visit inference in regression discontinuity designs with an eye towards generalizable procedures with finite-sample guarantees.

Linear estimators for RDD Recall that local linear regression estimates τ_c as follows. For a bandwidth $h_n > 0$ and a symmetric weighting function $K(\cdot)$, use

$$\begin{aligned}\hat{\tau}_c = \text{argmin} \Bigg\{ \sum_{i=1}^{n} K\left(\frac{|Z_i - c|}{h_n}\right) \\ \times \left(Y_i - a - \tau W_i - \beta_{(0)} (Z_i - c)_- - \beta_{(1)} (Z_i - c)_+\right)^2 \Bigg\},\end{aligned} \tag{6.2}$$

where the overall intercept a and and slope parameters $\beta_{(w)}$ are nuisance parameters. Popular choices for the weighting function $K(x)$ include the window kernel $K(x) = 1(\{|x| \leq 1\})$, or the triangular kernel $K(x) = (1 - |x|)_+$.

Now, when studying the local linear estimator (6.2) last time, we noted that we can write this estimator as

$$\hat{\tau}_c(\gamma) = \sum_{i=1}^{n} \gamma_i Y_i. \tag{6.3}$$

for some weights γ_i that only depend on the running variable Z_i. In the previous lecture, we wrote down a closed form expression for γ_i for the window kernel $K(x) = 1(\{|x| \leq 1\})$; however, from basic properties of least squares regression we see that such a linear representation is always available. And interestingly, despite the definition (6.2) of the local linear estimator, it turns out that we didn't make much use of this definition in studying $\hat{\tau}_c$. Instead, for our formal discussion, we just used general properties of linear estimators of the form (6.3).[1]

More specifically, our analysis of local linear regression only made use of the following fact that pertains to all linear estimators. For simplicity, let's work with homoskedatic and Gaussian errors, such that $Y_i(w) = \mu_{(w)}(Z_i) + \varepsilon_i(w)$ with $\varepsilon_i(w) \,\big|\, Z_i \sim \mathcal{N}(0, \sigma^2)$. Then, provided the weights γ_i are only functions of the Z_i, we have

$$\begin{aligned}\hat{\tau}_c(\gamma) \,\big|\, \{Z_1, ..., Z_n\} \;&\sim\; \mathcal{N}\left(\hat{\tau}_c^*(\gamma), \sigma^2 \|\gamma\|_2^2\right), \\ \hat{\tau}_c^*(\gamma) &= \sum_{i=1}^{n} \gamma_i \mu_{W_i}(Z_i),\end{aligned} \tag{6.4}$$

where $W_i = 1(\{Z_i \geq c\})$. Thus, we immediately see that any linear estimator as in (6.3) will be an accurate estimator for τ_c provided we can guarantee that $\hat{\tau}_c^*(\gamma) \approx \tau_c$ and $\|\gamma\|_2^2$ is small.

[1] There's a somewhat unfortunate naming collision here: When we say that local linear regression (6.2) is a linear estimator (6.3), we're using the qualifier linear twice with two different meanings.

Minimax linear estimation Motivated by this observation, it's natural to ask: If the salient fact about local linear regression (6.2) is that we can write it as an linear estimator of the form (6.3), then is local linear regression the best estimator in this class? As we'll see below, the answer is no; however, the best estimator of the form (6.3) can readily be derived in practice via numerical convex optimization.

As noted in (6.4), the conditional variance of any linear estimator can directly be observed: it's just $\sigma^2 \|\gamma\|_2^2$ (again, for simplicity, we're working with homoskedatic errors for most of today). In contrast, the bias of linear estimators depends on the unknown functions $\mu_{(w)}(z)$, and so cannot be observed:

$$\text{Bias}\left(\hat{\tau}_c(\gamma) \,\middle|\, \{Z_1, ..., Z_n\}\right) = \sum_{i=1}^{n} \gamma_i \mu_{W_i}(Z_i) - \left(\mu_{(1)}(c) - \mu_{(0)}(c)\right). \tag{6.5}$$

However, although, this bias is unknown, it can still readily be bounded given smoothness assumptions on the $\mu_{(w)}(z)$.

As in last lecture, consider the widely used smoothness assumption according to which the conditional response functions have bounded second derivatives, $|\mu''_{(w)}(z)| \leq B$. Then[2]

$$\begin{aligned} &\left|\text{Bias}\left(\hat{\tau}_c(\gamma) \,\middle|\, \{Z_1, ..., Z_n\}\right)\right| \leq I_B(\gamma) \\ &I_B(\gamma) = \sup\left\{\sum_{i=1}^{n} \gamma_i \mu_{W_i}(Z_i) - \left(\mu_{(1)}(c) - \mu_{(0)}(c)\right) : \left|\mu''_{(w)}(z)\right| \leq B\right\}. \end{aligned} \tag{6.6}$$

Now, recall that the mean-squared error of an estimator is just the sum of its variance and squared bias. Because the variance term $\sigma^2 \|\gamma\|_2^2$ doesn't depend on the conditional response functions, we thus see that the worst-case mean squared error of any linear estimator over all problems with $|\mu''_{(w)}(z)| \leq B$ is just the sum of its variance and worst-case bias squared, i.e.,

$$\text{MSE}\left(\hat{\tau}_c(\gamma) \,\middle|\, \{Z_1, ..., Z_n\}\right) \leq \sigma^2 \|\gamma\|_2^2 + I_B^2(\gamma), \tag{6.7}$$

with equality at any function that attains the worst-case bias (6.6).

It follows that, under an assumption that $|\mu''_{(w)}(z)| \leq B$ and conditionally on $\{Z_1, ..., Z_n\}$, the minimax linear estimator of the form (6.3) is the one that minimizes (6.7):

$$\hat{\tau}_c\left(\gamma^B\right) = \sum_{i=1}^{n} \gamma_i^B Y_i, \quad \gamma^B = \text{argmin}\left\{\sigma^2 \|\gamma\|_2^2 + I_B^2(\gamma)\right\}. \tag{6.8}$$

[2]There is no need for an absolute value inside the sup-term used to define $I_B(\gamma)$ because the class of twice differentiable functions is symmetric around zero. This fact will prove to be useful down the road.

One can check numerically that the weights implied by local linear regression do not solve this optimization problem, and so the estimator (6.8) dominates local linear regression in terms of worst-case MSE.

Deriving the minimax linear weights Of course, the estimator (6.8) is not of much use unless we can solve for the weights γ_i^B in practice. Luckily, we can do so via routine quadratic programming. To do so, it is helpful to write

$$\mu_{(w)}(z) = a_{(w)} + \beta_{(w)}(z - c) + \rho_{(w)}(z), \tag{6.9}$$

where $\rho_{(w)}(z)$ is a function with $\rho_{(w)}(c) = \rho'_{(w)}(c) = 0$ and whose second derivative is bounded by B; given this representation $\tau_c = a_{(1)} - a_{(0)}$.

Now, the first thing to note in (6.9) is that the coefficients $a_{(w)}$ and $\beta_{(w)}$ are unrestricted. Thus, unless the weights γ_i account for them exactly, such that

$$\sum_{i=1}^n \gamma_i W_i = 1, \quad \sum_{i=1}^n \gamma_i = 0, \quad \sum_{i=1}^n \gamma_i (Z_i - c)_+ = 0, \quad \sum_{i=1}^n \gamma_i (Z_i - c)_- = 0,$$

we can choose $a_{(w)}$ and $\beta_{(w)}$ to make the bias of $\hat{\tau}_c(\gamma)$ arbitrarily bad (i.e., $I_B(\gamma) = \infty$). Meanwhile, once we enforce these constraints, it only remains to bound the bias due to $\rho_{(w)}(z)$, and so we can re-write (6.8) as

$$\begin{aligned}
\left\{\gamma^B,\, t\right\} = \text{argmin} \quad & \sigma^2 \left\|\gamma\right\|_2^2 + B^2 t^2 \\
\text{subject to:} \quad & \sum_{i=1}^n \gamma_i W_i \rho_{(1)}(Z_i) + \sum_{i=1}^n \gamma_i (1 - W_i) \rho_{(0)}(Z_i) \leq t \\
& \qquad \text{for all } \rho_{(w)}(\cdot) \text{ with } \rho_{(w)}(c) = \rho'_{(w)}(c) = 0 \\
& \qquad\qquad \text{and } \left|\rho''_{(w)}(z)\right| \leq 1 \\
& \sum_{i=1}^n \gamma_i W_i = 1, \quad \sum_{i=1}^n \gamma_i = 0, \\
& \sum_{i=1}^n \gamma_i W_i (Z_i - c) = 0, \quad \sum_{i=1}^n \gamma_i (1 - W_i)(Z_i - c) = 0.
\end{aligned} \tag{6.10}$$

Given this form, the optimization should hopefully look like a tractable one. And in fact it is: The problem simplifies once we take its dual, and it can then be well approximated by a finite-dimensional quadratic program where we use a discrete approximation to the set of functions with second derivative bounded by 1. For details, see Section II.B of Imbens and Wager [2019].

Inference with linear estimators The above discussion suggests that using $\hat{\tau}_c\left(\gamma^B\right) = \sum_{i=1}^n \gamma_i^B Y_i$ with weights chosen via (6.10) results in a good point estimate for for τ_c if all we know is that $|\mu''_{(w)}(z)| \leq B$. In particular, under this assumption and conditionally on $\{Z_1, ..., Z_n\}$, it attains minimax mean-squared error among all linear estimators. Because local linear regression is also a linear estimator, we thus find that $\hat{\tau}_c\left(\gamma^B\right)$ dominates local linear regression in a minimax sense.[3]

If we want to use $\hat{\tau}_c\left(\gamma^B\right)$ in practice, though, it's important to be able to also provide confidence intervals for τ_c. And, since $\hat{\tau}_c\left(\gamma^B\right)$ balances out bias and variance by construction, we should not expect our estimator to be variance dominated—and any inferential procedure should account for bias.

To this end, recall (6.4), whereby conditionally on $\{Z_1, ..., Z_n\}$, the errors of our estimator, $\text{err} := \hat{\tau}_c - \tau_c$, are distributed as

$$\text{err} \,\big|\, \{Z_1, ..., Z_n\} \sim \mathcal{N}\left(\text{bias}, \sigma^2 \left\|\gamma^B\right\|_2^2\right). \tag{6.11}$$

Furthermore, the optimization problem (6.10) yields as a by-product an upper bound for the bias in terms of the optimization variable t, namely $|\text{bias}| \leq Bt$.

We can then use these facts to build confidence intervals as follows. Because the Gaussian distribution is unimodal,

$$\mathbb{P}\left[|\text{err}| \geq \zeta\right] \leq \mathbb{P}\left[\left|Bt + \sigma \left\|\gamma^B\right\|_2 S\right| \geq \zeta\right], \quad S \sim \mathcal{N}(0, 1). \tag{6.12}$$

Thus, we obtain level-α confidence intervals as follows:

$$\begin{aligned}
&\mathbb{P}\left[\tau_c \in \mathcal{I}_\alpha \,\big|\, \{Z_1, ..., Z_n\}\right] \geq 1 - \alpha, \\
&\mathcal{I}_\alpha = \left(\hat{\tau}_c\left(\gamma^B\right) - \zeta_\alpha^B,\ \hat{\tau}_c\left(\gamma^B\right) + \zeta_\alpha^B\right), \\
&\zeta_\alpha^B = \inf\left\{\zeta : \mathbb{P}\left[\left|Bt + \sigma \left\|\gamma^B\right\|_2 S\right| > \zeta\right] \leq \alpha, \quad S \sim \mathcal{N}(0, 1)\right\}.
\end{aligned} \tag{6.13}$$

In addition to formally accounting for bias, note that these intervals hold conditionally on Z_i, and so hold without any distributional assumptions on the running variable. This is useful when considering regression discontinuities in non-standard settings.

Example: Discrete running variable A first example of the usefulness of having conditional-on-Z_i guarantees is when the running variable Z_i has discrete support. In this case, the regression discontinuity parameter τ_c is in general not point-identified under only the assumption $|\mu''_{(w)}(z)| \leq B$ because

[3] Of course, one also needs to verify that $\hat{\tau}_c\left(\gamma^B\right)$ is not the same as local linear regression; this can be done numerically.

there may not be any data arbitrarily close to the boundary.[4] And, without point identification, any approach to inference that relies on asymptotics with specific rates of convergence for $\hat{\tau}_c$ as discussed in the previous lecture clearly is not applicable.

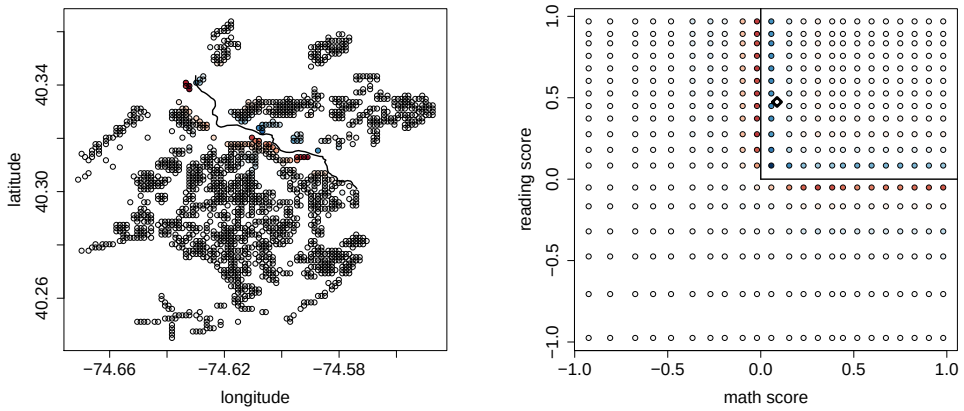
In contrast, in our case, the fact that Z_i may have discrete support changes nothing. The confidence intervals (6.13) have coverage conditionally on $\{Z_1, ..., Z_n\}$, and the empirical support $\{Z_1, ..., Z_n\}$ of the running variable is always discrete, so the question of whether the Z_i have a density in the population is irrelevant when working with (6.13). The relevance of a discrete Z_i only comes up asymptotically: If Z_i has a continuous density, then the confidence intervals (6.13) will shrink asymptotically at the optimal rate discussed in last lecture, namely $n^{-2/5}$. Conversely, if the Z_i has discrete support, the length of the confidence intervals will not go to 0; rather, we end up in a partial identification problem.

Example: Multivariate running variable So far, we have focused on regression discontinuity designs where treatment is determined by a single threshold: $W_i = 1(\{Z_i \geq c\})$ for some $Z_i \in \mathbb{R}$. However, the ideas discussed here apply in considerably more generality: One can let the running variable $Z_i \in \mathbb{R}^k$ be multivariate, and the treatment region be generic, i.e., $W_i = 1(\{Z_i \in \mathcal{A}\})$ for some set $\mathcal{A} \subset \mathbb{R}^k$. For example, in an educational setting, $Z_i \in \mathbb{R}^3$ could measure test results in 3 separate subjects, and $\mathcal{A}$ could denote the set of overall "passing" results given by, e.g., 2 out of 3 tests clearing a pass/fail cutoff. Or in a geographic regression discontinuity design, $Z_i \in \mathbb{R}^2$ could denote the location of one's household and $\mathcal{A}$ the boundary of some administrative region that deployed a specific policy.

The crux of a regression discontinuity design is that we seek to identify causal effects via sharp changes to an existing treatment assignment policy; and we can then apply the same reasoning as before to identify treatment effects along the boundary of the treatment region $\mathcal{A}$. That being said, while the extension of regression discontinuity designs to general multivariate settings is conceptually straight-forward, the methodological extensions require some more care. In particular, it is not always clear what the best way is to generalize local linear regression to a geographic regression discontinuity design.[5]

[4] When Z_i has a discrete distribution, the definition of τ_c via (6.1) needs careful interpretation—as we need to be able to talk about $\mu_{(w)}(z)$ at values of z that do not belong to the support of the running variable. All guarantees provided here hold if we define $\mu_{(w)}(z)$ outside of the support of z to be an arbitrary function that interpolates between the support points of z while satisfying $|\mu''_{(w)}(z)| \leq B$.

[5] When working with geographic regression discontinuities, some authors have tried to



The minimax linear approach, however, extends direction to a multivariate setting. When working with a multivariate running variable, one can essentially write down (6.10) verbatim, and interpret the resulting weighted estimator similarly to before. The resulting optimization problem is harder (one needs to optimize over multivariate non-parametric functions with bounded curvature), but nothing changes conceptually. The above figures above illustrate two weighting functions derived using this approach—once in a geographic setting, and once in an educational setting (a student needed to pass two tests to avoid a remedial program). Red points denote positive values of γ_i whereas blue dots denote negative values of γ_i; the strength of the color denotes magnitude of the weight.

Beyond homoskedaticity So far, we have focused on estimation and inference in the case where the noise $\varepsilon_i = Y_i - \mu_{(W_i)}(Z_i)$ was Gaussian with a known constant variance parameter σ^2. In practice, of course, neither of these assumptions is likely to hold. The upshot is that the conditional Gaussianity result (6.11) no longer holds exactly; rather, we need to invoke a central limit theorem to argue that

$$\hat{\tau}_c(\gamma) \,\big|\, \{Z_1, \ldots, Z_n\} \;\approx\; \mathcal{N}\left(\hat{\tau}_c^*(\gamma), \sum_{i=1}^{n} \gamma_i^2 \operatorname{Var}\left[Y_i \,\big|\, Z_i, W_i\right]\right). \tag{6.14}$$

However, provided we're willing to make assumptions under which the Gaussian approximation above is valid, we can still proceed as above to get confidence

collapse the problem by only considering a univariate running variable that codes distance to the boundary of $\mathcal{A}$. Such an approach, however, is sub-optimal from a statistical point of view as it throws away relevant information.

intervals. Meanwhile, we can (conservatively) estimate the conditional variance in (6.14) via

$$\widehat{V}_n = \sum_{i=1}^{n} \gamma_i^2 \left(Y_i - \hat{\mu}_{(W_i)}(Z_i)\right)^2, \tag{6.15}$$

where, e.g., $\hat{\mu}_{(W_i)}(Z_i)$ is derived via local linear regression; note that this bound is conservative if $\hat{\mu}_{(W_i)}(Z_i)$ is misspecified, since then the misspecifiaction error will inflate the residuals.

That being said, one should emphasize that the estimator (6.8) is only minimax under homoskedastic errors with variance σ^2; if we really wanted to be minimax under heteroskedasticity then we'd need to use per-parameter variances σ_i^2 in (6.10). Thus, one could argue that an analyst who uses the estimator (6.8) but builds confidence intervals via (6.14) and (6.15) is using an oversimplified homoskedastic model to motivate a good estimator, but then out of caution and rigor uses confidence intervals that allow for heteroskedasticity when building confidence intervals. This is generally a good idea, and in fact something that's quite common in practice (from a certain perspective, anyone who runs OLS for point estimation but then gets confidence intervals via the bootstrap is doing the same thing); however, it's important to be aware that one is making this choice.

Bibliographic notes The study of minimax linear estimators in problems of this type goes back to Donoho [1994], who showed to following result. Suppose that we want to estimate θ using a Gaussian random vector Y,

$$Y = Kv + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma I), \quad \theta = a \cdot v, \tag{6.16}$$

where the matrix K and vector a are know, but v is unknown. Suppose moreover that v is known to belong to a convex set $\mathcal{V}$. Then, there exists a linear estimator, i.e., an estimator of the form $\hat{\theta} = \sum_{i=1}^{n} \gamma_i Y_i$ whose risk is within a factor 1.25 of the minimax risk among all estimators (including non-linear ones), and the weights γ_i for the minimax linear estimator can be derived via convex optimization. From this perspective, the minimax RDD estimator (6.8) is a special case of the estimators studied by Donoho [1994],[6] and in fact his results imply that this estimator is nearly minimax among all estimators (not just linear ones).

In a first application of this principle to regression discontinuity designs, Armstrong and Kolesár [2018] study minimax linear estimation over a class

[6]Note that the class of functions with second derivative bounded by B is convex.

of function proposed by Sacks and Ylvisaker [1978] for which Taylor approximations around the cutoff c are nearly sharp. Our presentation today follows Imbens and Wager [2019], who consider numerical convex optimization for flexible inference in generic regression discontinuity designs. Finally, Kolesár and Rothe [2018] advocate worst-case bias measures of the form (6.6) as a way of avoiding asymptotics and providing credible confidence intervals in regression discontinuity designs with a discrete running variable.

When the running variable Z_i has a discrete distribution, τ_c is not point identified and so the problem of estimating this regression discontinuity problem is formally a partially identified problem. Although we do not pursue this perspective further here, we note that the bias-aware intervals (6.13) correspond exactly to a type of confidence interval for partially identified parameters proposed in Imbens and Manski [2004].

Lecture 7
Balancing Estimators

As emphasized in our discussion so far, the propensity score plays a key role in the estimation of average treatment effect estimation under unconfoundedness; we then considered inverse-propensity weighting (IPW) and augmented IPW (AIPW) as practical methods that leverage propensity score estimates. However, we did not discuss in detail how to estimate the propensity score such as to get the best possible statistical guarantees.

Our goal today is to revisit our discussion of IPW and AIPW estimators for the average treatment effect with an eye towards careful propensity estimation. In doing so, we'll build on insights from our discussion of regression discontinuity designs and use convex optimization to directly derive propensity weights with good finite sample properties.

Review: Why IPW works We're working under the potential outcomes model with IID samples $\{X_i,\, Y_i,\, W_i\} \in \mathcal{X} \times \mathbb{R} \times \{0,\, 1\}$, such that $Y_i = Y_i(W_i)$ for a pair of potential outcomes $\{Y_i(0),\, Y_i(1)\}$. Our goal is to estimate $\tau = \mu(1) - \mu(0)$, where $\mu(w) = \mathbb{E}\left[Y_i(w)\right]$. As usual, we assume

$$\text{Unconfoundedness:} \qquad \{Y_i(0),\, Y_i(1)\} \perp\!\!\!\perp W_i \,\big|\, X_i \tag{7.1}$$

$$\text{Overlap:} \qquad 0 < \eta \leq e(X_i) \leq 1 - \eta < 1, \tag{7.2}$$

where $e(x) = \mathbb{P}\left[W_i = 1 \,\middle|\, X_i = x\right]$ and η is some positive constant. Here, unconfoundedness is used to argue that controlling for X_i is sufficient for identifying the average treatment effect. Overlap implies that controlling for X_i is statistically practical.

For simplicity, today, we'll focus on estimating $\mu(1)$, since this allows for more compact notation while capturing the core conceptual issues. The inverse-propensity weighted estimator of $\mu(1)$ is

$$\hat{\mu}_{IPW}(1) = \frac{1}{n} \sum_{i=1}^{n} \frac{W_i Y_i}{\hat{e}(X_i)}, \tag{7.3}$$

where $\hat{e}(x)$ is an estimate of the propensity score $e(x)$. To prove consistency of IPW, we essentially argued as follows:

1. **Population balance.** Under unconfoundedness, $\mu(1) = \mathbb{E}\left[W_i Y_i / e(X_i)\right]$.

2. **Oracle estimator.** An oracle version $\hat{\mu}^*_{IPW}(1)$ of (7.3) with true propensity scores is unbiased; moreover, under overlap, it has finite variance.

3. **Feasible approximation.** Assume that $\hat{e}(X_i)$ also satisfies the overlap condition (7.2). Then, by Cauchy-Schwartz,

$$\begin{aligned} &|\hat{\mu}_{IPW}(1) - \hat{\mu}^*_{IPW}(1)| \\ &\qquad \leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{\hat{e}(X_i)} - \frac{1}{e(X_i)}\right)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(W_i Y_i\right)^2} \\ &\qquad \leq \frac{1}{\eta^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{e}(X_i) - e(X_i)\right)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(W_i Y_i\right)^2}. \end{aligned} \tag{7.4}$$

And, while this proof sketch obviously implies consistency, it also is not particularly sharp statistically. In particular, the Cauchy-Schwartz bound (7.4) clearly does not use any structure of the propensity estimates $\hat{e}(X_i)$, and makes rather crude use of the overlap assumption (7.4). In Lecture 3, we showed that augmented IPW could considerably improve the performance of IPW by using a regression adjustment and cross-fitting; however, the way we dealt with overlap still essentially amounts to the argument (7.4).[1]

Population vs. sample balance In order to understand how to design better variants of propensity score weighting, it is helpful to start by writing

[1]With AIPW, we found that whenever the nuisance component estimates converged fast enough, the estimation error in $\hat{e}(x)$ had a vanishing effect on the $1/\sqrt{n}$-scale and so any constants in (7.4) vanished into lower-order terms (and we did not discuss them much). However, if we want to get good behavior in regimes where errors in $\hat{e}(x)$ have a meaningful effect on the error of our average treatment effect (either because of finite sample effects or due to weaker guarantees on the rates of convergence of nuisance estimates made in Lecture 3), using a sharper argument than (7.4) is valuable.

the conditional response function $\mu_{(w)}(x)$ in terms of a basis expansion, i.e.,[2]

$$\mu_{(w)}(x) = \sum_{j=1}^{\infty} \beta_j(w)\, \psi_j(x) \tag{7.5}$$

for some pre-defined set of basis function $\psi_j(\cdot)$. Under reasonable regularity conditions, we then have

$$\mu(w) = \sum_{j=1}^{\infty} \beta_j(w)\, \mathbb{E}\left[\psi_j(X_i)\right]. \tag{7.6}$$

Given this notation, we can write down a revealing proof demonstrating that IPW is valid over the population. Under uncondoundedness, $Y_i = \mu_{(W_i)}(X_i) + \varepsilon_i$ with $\mathbb{E}\left[\varepsilon_i \,\middle|\, X_i,\, W_i\right] = 0$, and so (again under regularity conditions)

$$\begin{aligned} \mathbb{E}\left[\frac{W_i Y_i}{e(X_i)}\right] &= \mathbb{E}\left[\frac{W_i}{e(X_i)} \sum_{j=1}^{\infty} \beta_j(w)\, \psi_j(X_i)\right] \\ &= \sum_{j=1}^{\infty} \beta_j(w)\, \mathbb{E}\left[\frac{W_i \psi_j(X_i)}{e(X_i)}\right] = \sum_{j=1}^{\infty} \beta_j(w)\, \mathbb{E}\left[\psi_j(X_i)\right] = \mu(1) \end{aligned} \tag{7.7}$$

In other words, IPW works because weighting by $1/e(X_i)$ achieves population balance $\mathbb{E}\left[W_i \psi_j(X_i) \,/\, e(X_i)\right] = \mathbb{E}\left[\psi_j(X_i)\right]$ for all basis functions $j = 1,\, 2,\, ...$

This insight provides helpful guidance in how to think about good inverse-propensity weights. If the key property of the true propensity weights is that they achieve exact balance on the population, then a reasonable target to strive for with estimated propensity weights is that they achieve approximate balance in the sample:

$$\frac{1}{n} \sum_{i=1}^{n} \frac{W_i \psi_j(X_i)}{\hat{e}(X_i)} \approx \frac{1}{n} \sum_{i=1}^{n} \psi_j(X_i), \text{ for all } j = 1,\, 2,\, \ldots \tag{7.8}$$

The relevant notion of "$\approx$" is above depends on the setting; and we'll discuss several examples below. Overall, though, one should expect analyses of inverse-propensity weighting that go via the fundamental property (7.8) than the more indirect oracle approximation (7.4) to achieve sharper bounds.[3]

[2]The existence of such basis representations is well known in many contexts; for example, functions of bounded variation on a compact interval can be represented in terms of a Fourier series. Today, we'll not review when such representations are available; instead, we'll just work under the assumption that an appropriate series representation is given.

[3]In this context, it's interesting to recall our "aggregating" estimator from Lecture 2

Balancing loss functions for propensity estimation As a first example of learning propensity scores than emphasize finite-sample balance (7.8), consider a simple parametric specification: We assume a linear outcome model $\mu_{(w)}(x) = x \cdot \beta(w)$ and a logistic propensity model $e(x) = 1/(1 + e^{-x \cdot \theta})$. Because we have a linear outcome model, achieving sample balance just involves balancing the covariates X_i.

If we ask for exact balance (which is reasonable if we're in low dimensions) and want to use propensity scores given by a logistic model, then (7.8) becomes

$$\frac{1}{n}\sum_{i=1}^{n}\left(1 + e^{-X_i\hat{\theta}}\right) W_i X_i = \frac{1}{n}\sum_{i=1}^{n} X_i, \tag{7.9}$$

where the above equality is of two vectors in $\mathbb{R}^p$. The above condition may seem like a difficult non-linear equation; however, one can check that (7.9) is nothing but the KKT-condition for the following convex problem,[4]

$$\begin{aligned} \hat{\theta} &= \text{argmin}_\theta \left\{ \frac{1}{n}\sum_{i=1}^{n} \ell_\theta\left(X_i,\, Y_i,\, W_i\right) \right\}, \\ \ell_\theta\left(X_i,\, Y_i,\, W_i\right) &= W_i e^{-X_i\theta} + (1 - W_i) X_i \theta, \end{aligned} \tag{7.10}$$

and so has a unique solution that can be derived by Newton descent on (7.10).

The simple observation suggests that if we believe in a linear-logistic specification and want to use an IPW estimator, then we should learn the propensity model by minimizing the "balancing" loss function $\ell_\theta\left(X_i,\, Y_i,\, W_i\right)$ rather than by the usual method (i.e., logistic regression, meaning maximum likelihood in the logistic model). Maximum likelihood may be asymptotically optimal from the perspective of estimating the logistic regression parameters θ; but that's not what matters for the purpose of IPW estimation. What matters is that we fit a propensity model that satisfies (7.8), and for that purpose the loss function $\ell_\theta\left(X_i,\, Y_i,\, W_i\right)$ is better. In the homework, we'll study IPW with (7.10) further, and show that in the linear-logistic model this estimator performs well in terms of both robustness and asymptotic variance.

---

that applied when X_i had discrete support. Here, one obtains a representation (7.5) where $\psi_j(x)$ simply checks whether x is the j-th support point. Then, our aggregating estimator of the ATE corresponds to IPW with estimated propensity scores $\hat{e}(x)$ that account for the empirical treatment fraction for each value of x, and achieve *exact* sample balance (7.8).

[4]One interesting aspect of using (7.10) is that it requires us to learn different propensity models for getting IPW estimates of $\mu(0)$ and $\mu(1)$; and thus the resulting IPW estimator of the average treatment effect τ will use propensity estimates from two different propensity models. This may seem surprising, but is unavoidable if we want to achieve exact balance (since with have $2p$ balance conditions for estimating both $\mu(0)$ and $\mu(1)$, but the propensity model has only p free parameters θ).

ATE estimation with high-dimensional confounders As a second application of this balancing principle, consider the problem of ATE estimation in the high-dimensional linear model. We assume that unconfoundedness (7.1) holds, but only after controlling for covariates $X_i \in \mathbb{R}^p$ where p may be much larger than n (e.g., X_i may represent a patient's genome); moreover, as non-parametric analysis in high dimensions is generally intractable, we assume that $\mu_{(w)}(x) = x \cdot \beta(w)$ for some $\beta(w) \in \mathbb{R}^p$. In high dimensions, getting propensity score estimates that are stable enough for the argument (7.4) to go through is difficult, so the value of directly targeting balance as in (7.8) is particularly valuable.

Since we are in high dimensions, finding propensity weights that achieve exact balance as in (7.9) is not possible; the best we can hope for is approximate balance. With this in mind, we note that by Hölder's inequality,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\frac{W_i\mu_{(1)}(X_i)}{\hat{e}(X_i)} - \frac{1}{n}\sum_{i=1}^{n}\mu_{(1)}(X_i) \\
&\qquad = \left(\frac{1}{n}\sum_{i=1}^{n}\frac{W_iX_i}{\hat{e}(X_i)} - \frac{1}{n}\sum_{i=1}^{n}X_i\right)\beta(1) \qquad (7.11)\\
&\qquad \leq \left\|\frac{1}{n}\sum_{i=1}^{n}\frac{W_iX_i}{\hat{e}(X_i)} - \frac{1}{n}\sum_{i=1}^{n}X_i\right\|_\infty \|\beta(1)\|_1
\end{aligned}$$

This decomposition suggests a practical meaning for the "$\approx$" term in (7.8), namely that good inverse-propensity weights should achieve good worst-case imbalance across all features. But, although this decomposition gives some helpful insight, it is not easy to act on directly. In particular:

1. It is not obvious how to parametrize $\hat{e}(X_i)$ in order to achieve good sup-norm approximate balance as in (7.11).

2. The above bound is only meaningful with bounds on $\|\beta(1)\|_1$, but such bounds are not typically available (in high-dimensional statistics, it's common to assume that $\beta(1)$ should be sparse, but that still doesn't contain its 1-norm).

It turns out, however, that by combining ideas already covered in this class—namely optimizing for balance and augmented weighting estimators—with some basic lasso theory we can turn the insight (7.11) into a practical approach to high-dimensional inference about the ATE.

First, we note that the decomposition (7.11) makes no reference to the specific form of the inverse-propensity weights, so we can avoid the whole problem

of parametrization by optimizing for balance directly: For some well chosen $\zeta > 0$, let

$$\hat{\gamma} = \text{argmin}_\gamma \left\{ \frac{1}{n^2} \|\gamma\|_2^2 + \zeta \left\| \frac{1}{n} \sum_{i=1}^n (\gamma_i W_i - 1) X_i \right\|_\infty^2 : \gamma_i \geq 1 \right\}, \tag{7.12}$$

and then formally use "$1/\hat{e}(X_i) = \hat{\gamma}_i$" for inverse-propensity weighting. Assuming overlap (7.2), one can use sub-Gaussian concentration inequalities to check that the true inverse-propensity weights $e(X_i)$ satisfy

$$\frac{1}{n^2} \left\| \frac{1}{e(X_i)} \right\|_2^2 \leq \frac{\eta^{-2}}{n}, \quad \left\| \frac{1}{n} \sum_{i=1}^n (\gamma_i W_i - 1) X_i \right\|_\infty^2 = \mathcal{O}_P\left(\frac{\eta^{-2} \log(p)}{n} \right). \tag{7.13}$$

Thus, because (7.12) optimizes directly for the 2-norm of the γ as well as imbalance, we should expect its solution "$1/\hat{e}(X_i) = \hat{\gamma}_i$" to also satisfy the scaling bounds (7.13) for a good choice of the tuning parameter ζ, even if the implied propensity estimates $\hat{e}(X_i)$ may not be particularly good estimates of $e(X_i)$.

Second, we can fix the problematic dependence on $\|\beta(1)\|_1$ by augmenting our weighted estimator with a high-dimensional regression adjustment. Specifically, consider "augmented balancing" estimators of the form

$$\hat{\mu}_{AB}(1) = \frac{1}{n} \sum_{i=1}^n \hat{\mu}_{(1)}(X_i) + \hat{\gamma}_i W_i \left(Y_i - \hat{\mu}_{(1)}(X_i) \right), \quad \hat{\mu}_{(1)}(X_i) = X_i \hat{\beta}(1). \tag{7.14}$$

Then, emulating the argument in (7.11), we can check that

$$\begin{aligned} \hat{\mu}_{AB}(1) = & \underbrace{\frac{1}{n} \sum_{i=1}^n X_i \beta(1)}_{\text{sample avg. of } \mu_{(1)}(X_i)} + \underbrace{\frac{1}{n} \sum_{i=1}^n \hat{\gamma}_i W_i \left(Y_i - X_i \beta(1) \right)}_{\text{mean-zero noise term}} \\ & + \underbrace{\left(\frac{1}{n} \left(1 - \hat{\gamma}_i W_i \right) X_i \right) \left(\hat{\beta}(1) - \beta(1) \right)}_{\text{bias} \leq \left\| \frac{1}{n} (1 - \hat{\gamma}_i W_i) X_i \right\|_\infty \left\| \hat{\beta}(1) - \beta(1) \right\|_1}. \end{aligned} \tag{7.15}$$

Thus, much like with AIPW, we find that in augmented balancing the weights $\hat{\gamma}_i$ only need to correct for the regression error $\hat{\beta}(1) - \beta(1)$ rather than the full signal $\beta(1)$.

The reason the decomposition (7.15) matters is that, in high dimensional regression, there are many situations where we know how to get strong bounds

on $||\hat{\beta}(1) - \beta(1)||_1$. In particular, it is well-known that given sparsity $\|\beta(1)\|_0 \leq k$ and under restricted eigenvalue conditions, the lasso can achieve 1-norm error [e.g., Negahban, Ravikumar, Wainwright, and Yu, 2012]

$$\left\|\hat{\beta}(1) - \beta(1)\right\|_1 = \mathcal{O}_P\left(k\sqrt{\frac{\log(p)}{n}}\right). \tag{7.16}$$

We are now ready to put all the pieces together. Recall that we are in a high-dimensional linear setting as described above; and furthermore assume that conditions on X_i are satisfied so that (7.16) holds, and that $\beta(1)$ is k-sparse with $k \ll \sqrt{n}/\log(p)$. This sparsity condition is standard when proving resulting about high-dimensional inference [Javanmard and Montanari, 2014, Zhang and Zhang, 2014]. Then:

1. Start by running a lasso on the treated units. Given our sparsity condition $k \ll \sqrt{n}/\log(p)$ and (7.16), we find that the 1-norm error of $\hat{\beta}(1)$ decays as $o_P(1/\sqrt{\log(p)})$.

2. Fit weights $\hat{\gamma}$ as in (7.12). By the argument (7.13), we expect the infinity-norm imbalance to be of order $\mathcal{O}_P(\sqrt{\log(p)/n})$.

3. Estimate $\mu_{AB}(1)$ via (7.14). Plugging our two bounds from above into (7.15), we find that

$$\hat{\mu}_{AB}(1) - \underbrace{\frac{1}{n}\sum_{i=1}^{n} X_i\beta(1)}_{\text{sample avg. of } \mu_{(1)}(X_i)} = \underbrace{\frac{1}{n}\sum_{i=1}^{n} \hat{\gamma}_i W_i \left(Y_i - X_i\beta(1)\right)}_{\text{noise term}} + o_P\left(\frac{1}{\sqrt{n}}\right), \tag{7.17}$$

   where the noise term is of scale $\mathcal{O}_P(1/\sqrt{n})$.

Finally, we can use (7.17) for inference about the sample average of $\mu_{(1)}(X_i)$ by verifying the following self-normalized central limit theorem:

$$\frac{\sum_{i=1}^{n} \hat{\gamma}_i W_i \left(Y_i - X_i\beta(1)\right)}{\sqrt{\sum_{i=1}^{n} \hat{\gamma}_i^2 W_i^2 \left(Y_i - X_i\beta(1)\right)^2}} \Rightarrow \mathcal{N}\left(0,\, 1\right). \tag{7.18}$$

Wrapping up, we note that the argument used here has been a little bit heuristic (but it can be made rigorous), and the conclusion of (7.18) may have been slightly weaker than expected, as we only got confidence intervals for the sample average of $\mu_{(1)}(X_i)$, namely $\bar{\mu}(1) := n^{-1}\sum_{i=1}^{n} X_i\beta(1)$ rather than $\mu(1)$ itself.

Getting results about $\mu(1)$ would require showing convergence of the $\hat{\gamma}_i$, which we have avoided doing here.

That being said, we emphasize that the simple principle of balancing (7.8) enabled us to design a powerful approach to average treatment effect estimation in high dimensions using a mix of elementary ideas and well known facts about the lasso. Moreover, the high-level sparsity conditions required by our argument are in line with the usual conditions required for high-dimensional inference [Javanmard and Montanari, 2014, Zhang and Zhang, 2014].[5] This highlights the promise of balancing as a general principle for designing good average treatment effect estimators in new settings.

Closing thoughts Today, we talking about "balancing" as a fundamental explanation for why inverse-propensity weighting works, and as a principle for designing inverse-propensity weights that are better than just plugging in estimates $\hat{e}(X_i)$ obtained from off-the-shelf predictive methods. We also surveyed two applications of this idea: The design of covariate-balancing loss functions for improved estimation of parametric propensity models, and approximate balance as a guiding principle for high-dimensional ATE estimation.

What we did not do today is to consider balancing estimators as an alternative to AIPW in general non-parametric settings as considered in Lecture 3. Such results are available, but go beyond the scope of this class. As one example, Hirshberg and Wager [2017] consider augmented balancing estimators, but with weights chosen to balance a non-parametric class of functions. They find that, in considerable generality, the resulting estimator is semiparametrically efficient, and that the implied propensity scores that arise from solving for balance are universally consistent for the true propensity scores. Futhermore, they show that the conditions required for efficiency are in general competitive (although non-overlapping) with those required for AIPW, and that balancing—as expected—helps a great deal with poor overlap. Instead of requiring a strict overlap conditions as in (7.2), augmented balancing estimators work under the minimal condition required for the semiparametric efficient variance to exist, i.e., $\mathbb{E}\left[1/e(X_i)\right] < \infty$ and $\mathbb{E}\left[1/(1-e(X_i))\right] < \infty$.

More broadly, similar balancing phenomena play a key role in other refined average treatment effect estimators that attain efficiency under more general conditions than we obtained with our generic plug-in and cross-fit argument for AIPW in Lecture 3 [Kennedy, 2020, Newey and Robins, 2018].

[5] In fact, digging deeper, one can see that that the augmented balancing estimator discussed here and the debiased lasso of Javanmard and Montanari [2014] are two instantiations of exactly the same idea; see Section 3.1 of Athey, Imbens, and Wager [2018] for a further discussion.

Bibliographic notes The key role of covariate balance for average treatment effect estimation under unconfoundedness has long been recognized, and a standard operation procedure when working with any weighted or matching-type estimators is to use balance as a goodness of fit check [Imbens and Rubin, 2015]. For example, after fitting a propensity model by logistic regression, one could check that the induced propensity weights satisfy a sample balance condition of the type (7.8) with reasonable accuracy. If the balance condition is not satisfied, one could try fitting a different (better) propensity model.

The idea of using covariate balance as an idea to guide propensity estimation (rather than simply as a post-hoc sanity check) is more recent. Early proposals from different communities include Graham, de Xavier Pinto, and Egel [2012] Hainmueller [2012], Imai and Ratkovic [2014] and Zubizarreta [2015]. A unifying perspective on these methods via covariate-balancing loss functions is provided by Zhao [2019]. Meanwhile, Athey, Imbens, and Wager [2018] show that augmented balancing estimators can be used for ATE estimation in high dimensions, while Kallus [2016] considers a large class of non-parametric balancing estimators.

Finally, one should note that that the principles behind balanced estimation apply much more broadly than simply to average treatment effect estimation, and can in fact be used to estimate any linear functional with a well-behaved Riesz representer, i.e., any functional θ that can be characterized as $\theta = \mathbb{E}\left[\gamma(X_i)Y_i\right]$ in terms of a well-behaved Riesz representer $\gamma(\cdot)$.[6] One example of such a functional effect is the average effect of an infinitesimal nudge to a continuous treatment (i.e., the average derivative of the conditional response function with respect to the treatment variable). Chernozhukov, Escanciano, Ichimura, Newey, and Robins [2016] and Chernozhukov, Newey, and Robins [2018b] use this idea to build a family of AIPW-like estimators for general functionals, while Hirshberg and Wager [2017] consider efficiency properties of balancing-type estimators in this setting.

[6]Note that, in the case of estimating $\mu(1)$, the Riesz representer is $W_i/e(X_i)$), and the balance condition (7.7) is the type of condition typically used to define a Riesz representer.

Lecture 8
Methods for Panel Data

In this class so far, we've mostly worked with independent and identically distributed data. In many settings, however, the data has more complex structure that needs to be taken into account both for modeling and inference. Today, we'll focus on a specific type of structure that arises with panel (or longitudinal) data: We have data for $i = 1, ..., n$ units across $t = 1, ..., T$ time periods, and want to use this data to assess the effect of an intervention that affects some units in some time periods.

A constant treatment effect model For now, we'll focus on the following simple sampling model. For all $i = 1, \ldots, n$ and $t = 1, \ldots, T$, we observe an outcome $Y_{it} \in \mathbb{R}$ and a treatment assignment $W_{it} \in \{0, 1\}$. Furthermore, we assume that the treatments and outcomes are associated via the following constant effect model,

$$Y_{it} = Y_{it}(0) + W_{it}\tau, \text{ for all } i = 1, \ldots, n, \ t = 1, \ldots, T, \tag{8.1}$$

where $Y_{it}(0)$ is interpreted as the potential outcome we would have observed for the i-th unit at time t had they not been treated, and τ is interpreted as a constant treatment effect. We then seek to estimate τ.

The reason we work with this simple model is that it will allow us to quickly survey a fairly broad set of approaches to estimation with panel data. However, one should note that the simple model (8.1) has two major implications:

- There is no treatment heterogeneity, i.e., the treatment affects all units the same way in all time periods, and
- There are no treatment dynamics, i.e., a unit's outcome at time t is only affected by the treatment they receive at time t.

The lack of heterogeneity is not a particularly realistic assumption, but may still be a reasonable working assumption for a first attempt at a new setting. As

we've found repeatedly so far, if we design an estimator that targets a constant treatment effect parameter but then apply it to a setting with treatment heterogeneity, we'll usually end up converging to a weighted[1] treatment effect—as will also be the case here.

The second implication, i.e., no dynamics, is more severe, and obviously not applicable in many settings. For example, in healthcare, if a doctor gets a patient to exercise more at time t, this will probably affect their health at times $t' > t$, and not just at time t. And there are no general guarantees that methods that ignore dynamics recover anything reasonable in the presence of dynamics. We'll revisit this issue when we talk about dynamic treatment policies and reinforcement learning a few weeks from now (at which time we'll properly account for it). Today, however, we'll focus on the simplified setting (8.1), if for no other reason than because it's a setting that has traditionally received a considerable amount of attention, is widely used in applied work, and leads to some interesting statistical questions.

The two-way model The most classical way of instantiating (8.1) is by specifying a two-way additive structure for $Y_{it}(0)$, such that[2]

$$Y_{it} = \alpha_i + \beta_t + W_{it}\tau + \varepsilon_{it}, \qquad \mathbb{E}\left[\varepsilon \,\middle|\, \alpha,\, \beta,\, W\right] = 0. \tag{8.2}$$

In other words, we assume that each unit and each time period have a distinctive offset (or fixed effect), and that any deviation from this two-way structure is due to noise.

The two-way model (8.2) is very restrictive. However, in some simple situations, it leads to perfectly reasonable point estimates for τ. As a particularly nice example, consider the case where we only have two time periods ($T = 2$), and some units never get treated ($W_{i\cdot} = (0,\, 0)$) while others start treatment in the second period ($W_{i\cdot} = (0,\, 1)$). Then, OLS in (8.2) has a closed-form solution,

$$\begin{aligned}\hat{\tau} = &\frac{1}{|\{i : W_{i2} = 1\}|} \sum_{\{i:W_{i2}=1\}} (Y_{i2} - Y_{i1}) \\ &- \frac{1}{|\{i : W_{i2} = 0\}|} \sum_{\{i:W_{i2}=0\}} (Y_{i2} - Y_{i1}),\end{aligned} \tag{8.3}$$

[1]Note, however, that the weights may not all be positive; see de Chaisemartin and D'Haultfoeuille [2018] for a further discussion.

[2]One thing that's left implicit in the model below is that the treatment assignments are strictly exogenous, and may not depend on history. In other words, the assumption $\mathbb{E}\left[\varepsilon \,\middle|\, \alpha,\, \beta,\, W\right] = 0$ is like a stronger alternative to unconfoundedness that's embedded in a model.

i.e., we compare after-minus-before differences in outcomes for exposed vs. unexposed units. This "difference-in-differences" estimator is one that we might have derived directly from first principles, without going through (8.2), and clearly measures a relevant causal effect if exposure W_{i2} is randomly assigned.

Similar difference-in-differences arguments apply naturally when we only have two units, one of which never gets treated, and the other of which starts treatment at some time $1 < t' < T$. This structure appears in one of the early landmark studies using two-way modeling by Card and Krueger [1994], who compared employment outcomes in New Jersey, which raised its minimum wage, to those in Pennsylvania, which didn't.

In contrast, two-way models of the form (8.2) can be harder to justify in situations where both n and T are large and W_{it} has a generic distribution. By analogy to the case with only two time periods, the estimator resulting from running OLS is this two-way layout is still sometimes referred to as difference-in-differences; however, it no longer has a simple closed form solution.

One virtue of (8.2) is that it has strong observable implications. Among non-treated (and similarly treated) units, all trends should be parallel—because units only differ by their initial offset α_i. If parallel trends are not seen to hold in the data, the two-way model should not be used.

Finally, whenever using the two-way model for inference, one should model the noise term ε_{it} as dependent within rows. As a simplifying assumption, one might take the noise across rows to be IID with some generic covariance $\text{Var}[\varepsilon_{i\cdot}] = \Sigma$; however, as emphasized by Bertrand, Duflo, and Mullainathan [2004], taking each cell of ε_{it} to be independent is hard to justify conceptually and leads to suspect conclusions in practice. As a sanity check, in the $T = 2$ case (8.3) where the two-way model is on its strongest footing, the natural variance estimator takes variances of the differences,

$$\widehat{\text{Var}}[\hat{\tau}] = \frac{\widehat{\text{Var}}\left[Y_{i2} - Y_{i1} \,\middle|\, W_{i2} = 1\right]}{|\{i : W_{i2} = 1\}|} + \frac{\widehat{\text{Var}}\left[Y_{i2} - Y_{i1} \,\middle|\, W_{i2} = 0\right]}{|\{i : W_{i2} = 0\}|}, \tag{8.4}$$

which in fact corresponds to inference that's robust to ε_{it} being correlated within rows. More generally, one could consider inference for (8.2) via a bootstrap or jackknife that samples rows of the panel (as opposed to individual cells).

Interactive panel models A natural generalization of (8.2) is to allow units to have richer "types" that aren't fully captured by a single offset parameter α_i, and instead to write

$$Y_{it} = A_{i\cdot} B'_{t\cdot} + W_{it}\tau + \varepsilon_{it}, \quad \mathbb{E}\left[\varepsilon \,\middle|\, A, B, W\right] = 0, \quad A \in \mathbb{R}^{n\times k}, B \in \mathbb{R}^{T\times k}, \tag{8.5}$$

for some rank parameter k. Equivalently, one has $Y_{it} = L_{it} + W_{it}\tau + \varepsilon_{it}$ for some rank-k matrix L. The specification (8.5) is considerably more general than (8.2), and in particular no longer forces parallel trends (for example, some units may have high baselines but flat trends, whereas others may have low baselines but rapidly rising trends).

One approach to working with the model (8.5) is synthetic controls [Abadie, Diamond, and Hainmueller, 2010] and synthetic difference-in-differences [Arkhangelsky, Athey, Hirshberg, Imbens, and Wager, 2018]. Suppose that only units in the bottom-right corner of the panel are treated, i.e., $W_{it} = 1\left(\{i > n_0, \text{ and } t > T_0\}\right)$ for some $1 \leq n_0 < n$ and $1 \leq T_0 < T$. One common example of this structure arises when we are evaluating the effect of some new policy; in this case, we have one unit ($i = n$) that switches from control to treatment at time $t = T_0 + 1$, while all other units receive control throughout.

The idea of synthetic controls is to artificially re-weight the unexposed units (i.e., with $W_{i\cdot} = 0$) so that their average trend matches the (unweighted) average trend up to time t_0,[3]

$$\sum_{i=1}^{n_0} \gamma_i Y_{it} \approx \alpha + \frac{1}{n - n_0} \sum_{i=n_0+1}^{n} Y_{it}, \quad t = 1, ..., T_0, \tag{8.6}$$

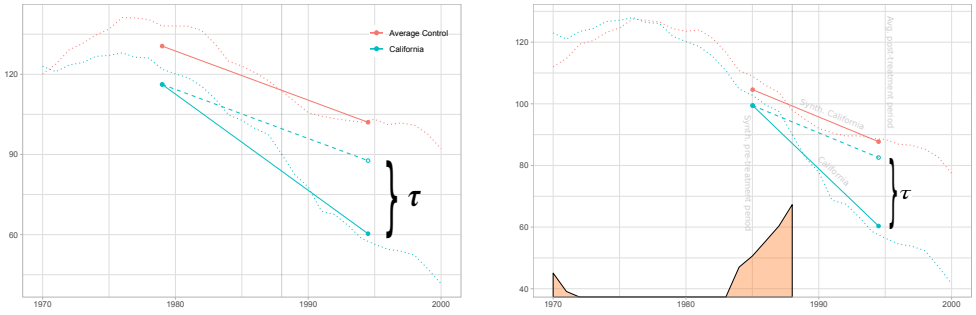
where α is an offset parameter analogous to a fixed effect. For example, one concrete choice of γ_i is to optimize squared-errors over the simplex:

$$\gamma = \text{argmin}_{\gamma', \alpha} \left\{ \left\| \sum_{i=1}^{n_0} \gamma_i' Y_{i(1:T_0)} - \frac{1}{n - n_0} \sum_{i=n_0+1}^{n} Y_{i(1:T_0)} - \alpha \right\|_2^2 : \sum_{i=1}^{n_0} \gamma_i' = 1, \, \gamma_i' \geq 0 \right\}. \tag{8.7}$$

The motivation behind this approach is that, if the weights γ succeed in creating parallel trends, then they should also be able to balance out the latent factors $A_{i\cdot}$. The upshot is that we can then estimate τ by weighted two-way regression,

$$\hat{\tau} = \text{argmin}_{\tau', \alpha', \beta'} \left\{ \sum_{i,t} \gamma_i \left(Y_{it} - \alpha_i - \beta_t - W_{it}\tau \right)^2 \right\}, \tag{8.8}$$

[3]The classical approach to synthetic controls following Abadie, Diamond, and Hainmueller [2010] does not allow for an offset, and instead seeks to match trajectories exactly. However, if we follow this weighting with a difference-in-differences regression as we do here, then there's we can allow for an offset.



difference-in-differences synthetic difference-in-differences

where we used the short-hand $\gamma_i = 1/(n - n_0)$ for $i > n_0$. Analogously to (8.3), this has a closed-form solution

$$\hat{\tau} = \frac{1}{n - n_0} \sum_{i=n_0+1}^{n} \left(\frac{1}{T - T_0} \sum_{t=T_0+1}^{T} Y_{it} - \frac{1}{T_0} \sum_{t=1}^{T_0} Y_{it} \right) - \sum_{i=1}^{n_0} \gamma_i \left(\frac{1}{T - T_0} \sum_{t=T_0+1}^{T} Y_{it} - \frac{1}{T_0} \sum_{t=1}^{T_0} Y_{it} \right). \tag{8.9}$$

Moreover, one can check that, under appropriate large-panel asymptotic (n_0, n, $T \to \infty$) this estimator is consistent and allows for asymptotically normal inference about τ in the low-rank specification (8.5); see Arkhangelsky et al. [2018] for details.

As an example of this idea consider the above figure, the goal of which is to illustrate the effect of a cigarette tax enacted in California in 1989 on smoking. We seek to identify this effect by comparing the prevalence of smoking in California to that of other states that did not enact a similar tax. The left panel compares the trend in California to the average trend in other states. Clearly these trends are not parallel, and so the model (8.2) is misspecified. The right panel, in contrast, shows how re-weighting the unexposed states using (8.7) lets us artificially create parallel trends; we can then estimate τ via (8.8). Note that, here, we also re-weighted the time periods $t = 1, ..., T_0$ via an analogue of (8.7) and used those weights in the regression (8.8); this is in general a good thing to do.

One should also note that synthetic controls are far from the only method that has been proposed for working with interactive fixed effects. Bonhomme and Manresa [2015] consider clustering the rows of the matrix Y via k-means and then fitting time-varying baseline models separately for each cluster, while

Athey, Bayati, Doudchenko, Imbens, and Khosravi [2017] propose estimating the low-rank baseline model directly via nuclear norm minimization. Finally, Bai [2009] studies asymptotics of a least-squares fit to (8.5) under specific assumptions on the factor matrices A and B. At the moment, however, flexible and general approaches for building uniformly valid confidence intervals for τ in the model (8.5) appear to be elusive.[4]

Identification via exchangeability Finally, a third approach to formalizing (8.1) is via "design-based" assumptions, whereby rows of the treatment assignment matrix $W_{i\cdot}$ are taken to be independent of baseline potential outcomes $Y_{i\cdot}$ conditionally on some observable event. One simple assumption of this type is

$$Y_{i\cdot}(0) \perp\!\!\!\perp W_{i\cdot} \,\big|\, S_i, \qquad S_i = \sum_{t=1}^{T} W_{it}, \tag{8.10}$$

i.e., that the distribution of treatment assignment for a unit is independent of their potential outcomes conditionally on the total number of time periods in which the unit receives treatment. This kind of assumption could be reasonable in, e.g., a healthcare application where Y_{it} corresponds to a health outcome and W_{it} represents receipt of preventive medical care—and we're worried by confounding due to unobserved health-seeking behavior (e.g., people who make sure to see their doctor regularly also take better care of themselves otherwise).

As noted by Arkhangelsky and Imbens [2019], one implication of (8.10) is that it enables unbiased estimation of τ via a wide variety of linear estimators, i.e., estimators of the form

$$\hat{\tau} = \sum_{i,\,t} \gamma_{it} Y_{it} \tag{8.11}$$

for some γ-matrix that only depends on the treatment assignment W_{it}. As a first step towards understanding good choices of γ, note that given (8.10) the rows of $Y_{it}(0)$ are exchangeable conditionally on S_i, and so

$$\mathbb{E}\left[\hat{\tau} \,\middle|\, W,\, \gamma\right] = \sum_{i,\,t} \gamma_{it} \mathbb{E}\left[Y_{it}(0) \,\middle|\, S_i\right] + \tau \sum_{i,\,t} \gamma_{it} W_{it}. \tag{8.12}$$

[4] In building confidence intervals, the approach of Arkhangelsky et al. [2018] heavily uses the fact that the treatment assignment region looks like a block $W_{it} = 1\left(\{i > n_0 \,\text{and}\, t > T_0\}\right)$, while the approach of Bai [2009] relies on strong-signal asymptotics that enable factor analysis to accurately recover A and B. Methods for inference on W that require special structure on neither W nor A and B would be of considerable interest.

Thus, the weighted estimator (8.11) is unbiased whenever $\sum_{i,\,t} \gamma_{it} W_{it} = 1$, and

$$\sum_{\{i:S_i=s\}} \gamma_{it} = 0 \text{ for all } t = 1, \ldots, \text{ and } s \in \mathcal{S}, \tag{8.13}$$

where $\mathcal{S}$ denotes the support of S_i. Then, one could try to pick γ by minimizing variance subject to these unbiasedness constraints,

$$\gamma = \text{argmin}_{\gamma'} \left\{ \sum_{i,\,t} \gamma_{it}^2 : \sum_{i,\,t} \gamma_{it} W_{it} = 1, \sum_{\{i:S_i=s\}} \gamma_{it} = 0 \text{ for all } s,\, t \right\}, \tag{8.14}$$

analogously to what we discussed with regression discontinuity designs in Lecture 6. This estimator will be unbiased whenever (8.10) holds and the optimization problem (8.14) is feasible—which it in general will be provided that W_{it} has non-trivial variation conditionally on S_i (i.e., the dataset must exhibit several different treatment assignment patterns $W_{i\cdot}$ with the same S_i).

Arkhangelsky and Imbens [2019] go further yet, and note that weighted estimators of the form (8.11) are also unbiased under the two-way model (8.2) provided the following constraints hold,

$$\sum_t \gamma_{it} = 0 \text{ for all } i = 1, ..., n, \quad \sum_i \gamma_{it} = 0 \text{ for all } t = 1, ..., T, \tag{8.15}$$

along with $\sum_{i,\,t} \gamma_{it} W_{it} = 1$. One can check unbiasedness by noting that the above equality constraints exactly cancel out the fixed effects α_i and β_t. Then, based on this observation, they propose a doubly robust estimator

$$\begin{aligned} \gamma = &\text{argmin}_{\gamma'} \sum_{i,\,t} \gamma_{it}^2 \\ &\text{subject to: } \sum_{i,\,t} \gamma_{it} W_{it} = 1, \quad \sum_{\{i:S_i=s\}} \gamma_{it} = 0 \text{ for all } s,\, t, \\ &\qquad \sum_t \gamma_{it} = 0 \text{ for all } i, \quad \sum_i \gamma_{it} = 0 \text{ for all } t. \end{aligned} \tag{8.16}$$

This estimator will be unbiased whenever (8.2) *or* (8.10) holds, provided the above optimization is feasible.

Bibliographic notes The study of panel (or longitudinal) data is a huge topic whose surface we've only scratched today. The list of topics we haven't covered today is too long to even attempt an enumeration. Arellano [2003] and Wooldridge [2010] present an overview of the area, and provide further references. In particular, there is a large body of work that focuses on ideas built around differencing (e.g., via generalizations of (8.3)) to identify causal effects; however, a discussion of such methods is beyond the scope of this class.

Lecture 9
Instrumental Variables Regression

Unconfoundedness is a powerful assumption, and plays a central role in many widely used approaches to identifying and estimating treatment effects in observational studies. In some applications, however, unconfoundedness is simply not plausible. For example, when studying the effect of prices on demand, it is unrealistic to assume that potential outcomes of demand (i.e., what demand would have been at given prices) is independent of what prices actually were. Instead, it's much more plausible to assume that prices and demand both respond to each other until a supply-demand equilibrium is reached. Today we'll introduce instrumental variables regression, which is a popular approach to measuring the effects of endogenous (i.e., not unconfounded) treatments.

A structural model In order to understand the principles behind instrumental variables regression, it is easiest to start with a simple constant treatment effects model. In the next lecture, we'll consider the behavior of instrumental variables methods in a general non-parametric setting with causal effects defined in terms of potential outcomes.

To this end, suppose we have outcome-treatment pairs (Y_i, W_i) satisfying a constant treatment effects model, such that

$$Y_i(w) = Y_i(0) + w\tau, \qquad Y_i = Y_i(W_i). \tag{9.1}$$

In our discussion so far, the next thing we've always done is to assume unconfoundedness, i.e., $\{Y_i(w)\} \perp\!\!\!\perp W_i$, which under model (9.1) reduces to $Y_i(0) \perp\!\!\!\perp W_i$. We can then re-write (9.1) as

$$Y_i = \alpha + W_i\tau + \varepsilon_i, \tag{9.2}$$

where $\alpha = \mathbb{E}\left[Y_i(0)\right]$, $\varepsilon_i = Y_i(0) - \mathbb{E}\left[Y_i(0)\right]$, and $\mathbb{E}\left[\varepsilon_i \,\middle|\, W_i\right] = 0$, and can consistently estimate τ by running OLS of Y_i on W_i.

Today, however, we're not going to assume unconfoundedness, and instead allow for a setting where $Y_i(0) \not\perp\!\!\!\perp W_i$. In this case, (9.2) still holds; however,

$\mathbb{E}\left[\varepsilon_i \,\middle|\, W_i\right] \neq 0$. In other words, (9.2) encodes a structural link between Y_i and W_i (really, it's just a way of writing (9.1) while hiding the potential outcomes), but it no longer captures a conditional expectation that can be analyzed using OLS. In particular, if we try estimating $\hat{\tau}_{OLS}$ by regressing Y_i on W_i, then in large samples we'll converge to

$$\begin{aligned} \tau_{OLS} &= \frac{\text{Cov}\left[Y_i,\, W_i\right]}{\text{Var}\left[W_i\right]} = \frac{\text{Cov}\left[\tau W_i + \varepsilon_i,\, W_i\right]}{\text{Var}\left[W_i\right]} \\ &= \tau + \frac{\text{Cov}\left[\varepsilon_i,\, W_i\right]}{\text{Var}\left[W_i\right]} \neq \tau. \end{aligned} \tag{9.3}$$

Note that, in the social sciences, it is quite common to write down linear relations of the form (9.2) that are intended to describe the structure of a system, but are not to be taken as a short-hand for linear regression. One the other hand, this is largely the opposite of standard practice in applied statistics where, when someone writes (9.2), they often don't mean anything else than that they intend to run a linear regression of Y_i on W_i.

Identification using instrumental variables In order to identify τ in model (9.2) without unconfoundedness, we need access to more data—and finding an instrument is one way to move forward. Qualitatively, an instrument is a variable Z_i that nudges the treatment level W_i but is uncorrelated with the noise term ε_i. For example, following an example of Angrist, Graddy, and Imbens [2000], consider a demand estimation problem where W_i is the price of fish and Y_i is demand. Then, one idea of an instrument Z_i could be to use weather conditions: Stormy weather makes it harder to fish (and thus raises prices), but presumably does not affect the demand curve.

Formally, we can add an instrument $Z_i \in \mathbb{R}$ to the structural model (9.2) as follows:

$$\begin{aligned} Y_i &= \alpha + W_i \tau + \varepsilon_i, \qquad \varepsilon_i \perp\!\!\!\perp Z_i \\ W_i &= Z_i \gamma + \eta_i. \end{aligned} \tag{9.4}$$

The fact that Z_i is uncorrelated with ε_i (or, in other words, that Z_i is exogenous) then implies that

$$\text{Cov}\left[Y_i,\, Z_i\right] = \text{Cov}\left[\tau W_i + \varepsilon_i,\, Z_i\right] = \tau \,\text{Cov}\left[W_i,\, Z_i\right], \tag{9.5}$$

and so the treatment effect parameter τ is identified as

$$\tau = \text{Cov}\left[Y_i,\, Z_i\right] \big/ \,\text{Cov}\left[W_i,\, Z_i\right]. \tag{9.6}$$

In other words, by bringing in an instrument, we've succeeded in identifying τ in (9.2) without unconfoundedness. The relation (9.6) also suggests a simple approach to estimating τ in terms of sample covariances, $\hat{\tau} = \widehat{\text{Cov}}\left[Y_i,\, Z_i\right] \,/\, \widehat{\text{Cov}}\left[W_i,\, Z_i\right]$.

In order for this identification strategy to work, the instrument Z_i needs to satisfy 3 key properties. First, Z_i must be *exogenous*, which here means $\varepsilon_i \perp\!\!\!\perp Z_i$; second, Z_i must be *relevant*, such that $\text{Cov}\left[W_i,\, Z_i\right] \neq 0$; finally, Z_i must satisfy the *exclusion restriction*, meaning that any effect of Z_i on Y_i must be mediated via W_i. Here, the exclusion restriction is baked in to the functional form (9.4). In the next lecture, we'll take a closer look at all these assumption in the context of a non-parametric specification.

Optimal instruments Above, we assumed that we had access to a single real-valued instrument Z_i, which essentially automatically lead us to the identification result (9.6). In practice, however, we may have access to many (potentially unstructured) candidate instruments Z_i: For example, when studying the effect of prices on demand for fish, we could consider storminess, year-to-year variation in the abundance of fish stock, and availability of imported fish as candidate instruments. This leads to the following more general specification,

$$Y_i = \tau W_i + \varepsilon_i, \quad \varepsilon_i \perp\!\!\!\perp Z_i, \quad Y_i,\, W_i \in \mathbb{R}, \quad Z_i \in \mathcal{Z}, \tag{9.7}$$

where $\mathcal{Z}$ may be, e.g., a high-dimensional space.

Because Z_i now takes values in a general space $\mathcal{Z}$, the statement (9.6) no longer makes sense. However, by the same argument as in (9.5), we see that given any function $w : \mathcal{Z} \to \mathbb{R}$ that maps Z_i to the real line, we have

$$\tau = \frac{\text{Cov}\left[Y_i,\, w(Z_i)\right]}{\text{Cov}\left[W_i,\, w(Z_i)\right]} \tag{9.8}$$

provided the denominator is non-zero (i.e., provided $w(Z_i)$ in fact "nudges" the treatment). In other words, if one has access to many valid instruments, the analyst is free to compress them into any univariate instrument of their choice.

Now, given the result (9.8), it's of course natural to ask what the optimal transformation $w(\cdot)$ is. To do so, note that the estimator suggested by (9.8),

$$\hat{\tau}_w = \frac{\widehat{\text{Cov}}\left[Y_i,\, w(Z_i)\right]}{\widehat{\text{Cov}}\left[W_i,\, w(Z_i)\right]} = \frac{\frac{1}{n}\sum_{i=1}^n \left(Y_i - \overline{Y}\right)\left(w(Z_i) - \overline{w(Z)}\right)}{\frac{1}{n}\sum_{i=1}^n \left(W_i - \overline{W}\right)\left(w(Z_i) - \overline{w(Z)}\right)} \tag{9.9}$$

with $\overline{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$, etc., is the solution to an estimating equation:

$$\frac{1}{n}\sum_{i=1}^n \left(w(Z_i) - \overline{w(Z)}\right)\left(Y_i - \overline{Y} - \hat{\tau}_w\left(W_i - \overline{W}\right)\right) = 0. \tag{9.10}$$

We can thus derive the asymptotic variance of $\hat{\tau}_w$ via general results about estimating equations, and find that[1]

$$n\left(\hat{\tau}_w - \tau\right) \Rightarrow \mathcal{N}\left(0,\, V_w\right), \quad V_w = \frac{\text{Var}\left[\varepsilon_i\right] \text{Var}\left[w(Z_i)\right]}{\text{Cov}\left[W_i,\, w(Z_i)\right]^2}, \tag{9.11}$$

where we note that $\text{Var}\left[\varepsilon_i \left(w(Z_i) - \mathbb{E}\left[w(Z_i)\right]\right)\right] = \text{Var}\left[\varepsilon_i\right] \text{Var}\left[w(Z_i)\right]$ by independence of Z_i and ε_i. Thus, the optimal instrument is the one that minimizes the limiting variance, i.e.,

$$w^*(\cdot) \in \text{argmax}_{w'} \left\{ \text{Cov}\left[W_i,\, w'(Z_i)\right]^2 \,/\, \text{Var}\left[w'(Z_i)\right] \right\}. \tag{9.12}$$

This is a well-known maximization problem, with solution $w^*(z) \propto \mathbb{E}\left[W_i \,\middle|\, Z_i = z\right]$. In other words, the optimal instrument $w^*(Z_i)$ is nothing but the best prediction of W_i from Z_i.

Cross-fitting and feasible estimation Given our above finding that the optimal instrument is the solution to a non-parametric prediction problem, $w^*(z) = \mathbb{E}\left[W_i \,\middle|\, Z_i = z\right]$, one might be tempted to apply the following two-stage strategy:

1. Fit a non-parametric first stage regression, resulting in estimate $\hat{w}(\cdot)$ of $\mathbb{E}\left[W_i \,\middle|\, Z_i = z\right]$, and then
2. Run (9.9) with $\hat{w}(\cdot)$ as an instrument.

This approach *almost* works, but may suffer from striking overfitting bias when the instrument is weak. The main problem is that, if $\hat{w}(Z_i)$ is fit on the training data, then we no longer have $\hat{w}(Z_i) \perp\!\!\!\perp \varepsilon_i$ (because $\hat{w}(Z_i)$ depends on W_i, which in turn is dependent on ε_i). This may seem like a subtle problem but, as pointed out by Bound, Jaeger, and Baker [1995], this may be a huge problem in practice; for example, they exhibit an example where the instrument Z_i is pure noise, yet the direct two-stage estimator $\hat{w}(Z_i)$ converges to a definite quantity (namely the simple regression coefficient $\text{OLS}(Y_i \sim W_i)$ which, because of lack of unconfoundedness, cannot be interpreted as a causal quantity).

Thankfully, however, we can again use cross-fitting to solve this problem. Specifically, we randomly split data into folds $k = 1, ..., K$ and, for each k, fit a regression $\hat{w}^{(-k)}(z)$ on all but the k-th fold. We then run

$$\hat{\tau} = \widehat{\text{Cov}}\left[Y_i,\, \hat{w}^{(-k(i))}(Z_i)\right] / \widehat{\text{Cov}}\left[W_i,\, \hat{w}^{(-k(i))}(Z_i)\right], \tag{9.13}$$

[1]Recall that if θ solves $\mathbb{E}\left[\psi_i(\theta)\right] = 0$ for some random function ψ_i and we estimate $\hat{\theta}$ via $\frac{1}{n}\sum_{i=1}^n \psi_i(\hat{\theta}) = 0$, then the non-parametric delta-method tells us that, in general, $\sqrt{n}(\hat{\theta} - \theta) \Rightarrow \mathcal{N}\left(0,\, V\right)$ with $V = \text{Var}\left[\psi_i(\theta)\right] / \mathbb{E}\left[\psi_i'(\theta)\right]^2$.

where $k(i)$ picks out the data fold containing the i-th observation. Now, by cross-fitting we directly see that $\hat{w}^{(-k(i))}(Z_i) \perp\!\!\!\perp \varepsilon_i$, and so this approach recovers a valid estimate of τ. In particular, if the regressions $\hat{w}^{(-k(i))}(z)$ are consistent for $\mathbb{E}\left[W_i \,\middle|\, Z_i = z\right]$ in mean-squared error, then the feasible estimator (9.13) is first-order equivalent to (9.9) with an optimal instrument.

Non-parametric instrumental variables regression One major assumption we've made today is that $Y_i = W_i\tau + \varepsilon_i$ as in (9.7), i.e., that the instrument acts linearly on Y_i. Next time, we'll talk about how to relax this assumption using potential outcomes notation. However, another generalization of (9.7) worth mentioning is what's commonly called the non-parametric instrumental variables problem,

$$Y_i = g(W_i) + \varepsilon_i, \quad Z_i \perp\!\!\!\perp \varepsilon_i, \quad Y_i, W_i \in \mathbb{R}, \quad Z_i \in \mathcal{Z}, \tag{9.14}$$

where $g(\cdot)$ is some generic smooth function we want to estimate. As before, because W_i is not independent of ε_i, we cannot learn $g(\cdot)$ by simply doing a (non-parametric) regression of Y_i on W_i, i.e., $g(w) \neq \mathbb{E}\left[Y_i \,\middle|\, W_i = w\right]$.

Instead, we should interpret (9.14) as a structural model that needs to be fit using the instrument. Because $Z_i \perp\!\!\!\perp \varepsilon_i$ and assuming that $\mathbb{E}\left[\varepsilon_i\right] = 0$, we can directly verify that

$$\begin{aligned} \mathbb{E}\left[Y_i \,\middle|\, Z_i = z\right] &= \mathbb{E}\left[g(W_i) + \varepsilon_i \,\middle|\, Z_i = z\right] \\ &= \mathbb{E}\left[g(W_i) \,\middle|\, Z_i = z\right] \\ &= \int_{\mathbb{R}} g(w) f\left(w \,\middle|\, z\right) \, dw, \end{aligned} \tag{9.15}$$

where $f(w \,|\, z)$ denotes the conditional density of W_i given $Z_i = z$. This relationship suggests a two-stage scheme for learning $g(\cdot)$, whereby we (1) fit a non-parametric model $\hat{f}(w \,|\, z)$ for the conditional density $f(w \,|\, z)$, preferably using cross-fitting, and (2) estimate $g(w)$ via a empirical minimization over a suitably chosen function class $\mathcal{G}$,

$$\hat{g}(\cdot) = \text{argmin}_{g \in \mathcal{G}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \int_{\mathbb{R}} g(w) \hat{f}^{(-k(i))}\left(w \,\middle|\, Z_i\right) \, dw \right)^2 \right\}. \tag{9.16}$$

In order to solve the inverse problem (9.16) in practice, one approach is to approximate $g(w)$ in terms of a basis expansion, $g_K(w) = \sum_{k=1}^{K} \beta_k \psi_k(w)$, where the $\psi_k(\cdot)$ are a set of pre-determined basis functions and $g_K(w)$ provides an

increasingly good approximation to $g(w)$ as K gets large. Then, (9.16) becomes

$$\begin{aligned} \hat{\beta} &= \text{argmin}_\beta \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \hat{m}^{(-k(i))}(Z_i) \cdot \beta \right)^2 \right\}, \text{ where} \\ \hat{m}^{(-k(i))}(Z_i) &= \int_{\mathbb{R}} \psi_k(w) \, \hat{f}^{(-k(i))} \left(w \,\middle|\, Z_i \right) \, dw \end{aligned} \tag{9.17}$$

can be interpreted as a multivariate cross-fit optimal instrument by analogy to (9.13). Conditions under which this type of approach yields a consistent estimate of $g(\cdot)$ are discussed in Newey and Powell [2003]. In general, however, one should note that solving the integral equation (9.15) is a difficult inverse problem, and so getting (9.17) to work in practice requires careful regularization (and, even so, one should expect rates of convergence to be slow).

Bibliographic notes The study of statistical estimation in simultaneous equation models (e.g., for joint modeling of prices and demand) has a long tradition in econometric; see, e.g., Haavelmo [1943] for an early reference. Imbens [2014] provides a review of this line of work aimed for statisticians, and also provides references to the recent literature. One should also note that (9.4) is an instance of a very simple structural equations model [Pearl, 2009]. We'll study graphical methods for working with much richer models of this type later in the class.

The literature on efficient estimation with instrumental variables goes back to Amemiya [1974], Chamberlain [1987], and others. The formulation of the efficient estimation problem in terms of non-parametric prediction of W_i in terms of Z_i is due to Newey [1990]; in particular, his results imply that the estimator (9.9) with $w(z) = \mathbb{E}\left[W_i \,\middle|\, Z_i = z\right]$ is efficient for τ in the model (9.7). Belloni, Chen, Chernozhukov, and Hansen [2012] propose estimating this first stage regression using the lasso [Hastie, Tibshirani, and Wainwright, 2015].

One question we've ignored today is the role of covariates for instrumental variables regression. Following our approach to unconfoundedness, one can extend (9.7) such that $\varepsilon_i \perp\!\!\!\perp Z_i \,\middle|\, X_i$, i.e., the instrument is only exogenous after conditioning on X_i, and we have a heterogeneous treatment effect function identified as $\tau(x) = \text{Cov}\left[Y_i, w(Z_i) \,\middle|\, X_i = x\right] \,/\, \text{Cov}\left[W_i, w(Z_i) \,\middle|\, X_i = x\right]$; see Abadie [2003] for a further discussion. Given this setting, one can then re-visit many of the questions we considered under unconfoundedness. For example, Chernozhukov, Escanciano, Ichimura, Newey, and Robins [2016] show how to build a doubly robust estimator of the average effect $\tau = \mathbb{E}\left[\tau(X)\right]$, and Athey, Tibshirani, and Wager [2019] propose a random forest estimator of $\tau(\cdot)$.

Lecture 10
Local Average Treatment Effects

Instrumental variables are commonly used to estimate the effect of an endogenous treatment. Last time, we discussed how IV methods can be used to estimate a treatment parameter in a structural model. In particular, we showed that in the following two-stage model

$$\begin{aligned} Y_i &= \alpha + W_i\tau + \varepsilon_i, \quad \varepsilon_i \perp\!\!\!\perp Z_i \\ W_i &= Z_i\gamma + \eta_i, \end{aligned} \tag{10.1}$$

the parameter τ can be identified via

$$\begin{aligned} \tau &= \text{Cov}\left[Y_i,\, Z_i\right] \big/ \text{Cov}\left[W_i,\, Z_i\right] \\ &= \frac{\mathbb{E}\left[Y_i \,\middle|\, Z_i = 1\right] - \mathbb{E}\left[Y_i \,\middle|\, Z_i = 0\right]}{\mathbb{E}\left[W_i \,\middle|\, Z_i = 1\right] - \mathbb{E}\left[W_i \,\middle|\, Z_i = 0\right]}, \end{aligned} \tag{10.2}$$

where the second expression is valid only when Z_i is binary. Furthermore, this representation also suggests a natural estimator for τ in terms of empirical covariances.

In general, however, the causal inference community is often skeptical of statistical targets that are only defined as parameters in a linear model. Thus, to further justify the relevance of IV methods to causal inference, today we'll revisit their behavior in the context of several concrete applications, e.g., non-compliance and demand modeling, with causal effects carefully defined in terms of potential outcomes. Our main finding will be that, in many settings, the natural IV estimator (10.2) targets a weighted treatment effect; furthermore, we'll also consider how to modify (10.2) to get at different weighted estimands. Today, we'll focus on questions of identification; the resulting estimation problems are closely related to the ones discussed last time.

Treatment effect estimation under non-compliance The simplest setting in which we can discuss non-parametric identification using instrumen-

tal variables is when estimating the effect of a binary treatment under non-compliance. Suppose, for example, that we've set up a randomized study to examine the effect of taking a drug to lower cholesterol. But although we randomly assigned treatment, some people don't obey the randomization: Some subjects given the drugs may fail to take them, while others who were assigned control may procure cholesterol lowering drugs on their own. In this case, we have

- An outcome $Y_i \in \mathbb{R}$, with the usual interpretation;
- The treatment $W_i \in \{0, 1\}$ that was actually received (i.e., did the subject take the drug), which is not random because of non-compliance; and
- The assigned treatment $Z_i \in \{0, 1\}$ which is random.

A popular way to analyze this type of data is using instrumental variables, where we interpret treatment assignment Z_i as an exogenous "nudge" on the treatment W_i that was actually received.[1]

If one believed in the structural model (10.1), then one could directly estimate τ via (10.2). In practice, however, we may not believe in the constant treatment effect assumption (10.1); e.g., one might ask whether people who comply with the treatment also would have responded differently to the treatment than others (maybe they chose to comply because they knew they'd benefit a lot from it).

A more careful approach starts by writing down potential outcomes. First, because W_i is non-random and may respond to Z_i, we need to have potential outcomes for the treatment variable in terms of the instrument, i.e., there are $\{W_i(0), W_i(1)\}$ such that $W_i = W_i(Z_i)$. Second, of course, we need to define potential outcomes for the outcome, which may in principle respond to both W_i and Z_i: we have $\{Y_i(w, z)\}_{w,z\in\{0,1\}}$ such that $Y_i = Y_i(W_i, Z_i)$. Given this notation, we now revisit our assumptions for what makes a valid instrument:

- *Exclusion restriction.* Treatment assignment only affects outcomes via receipt of treatment, i.e., $Y_i(w, z) = Y_i(w)$ for all w and z.
- *Exogeneity.* The treatment assignment is randomized, meaning that $\{Y_i(0), Y_i(1), W_i(0), W_i(1)\} \perp\!\!\!\perp Z_i$.

[1]Note that similar statistical patters also arise outside of clinical trials. For example, when studying the effect of military service on long-term income, one could write W_i for whether a person actually served in the military, and Z_i for the results of the draft lottery (i.e., did the government assign them to serve).

- *Relevance.* The treatment assignment affects receipt of treatment, meaning that $\mathbb{E}\left[W_i(1) - W_i(0)\right] \neq 0$.

Finally, we make one last assumption about how people respond to treatment. Defining each subject's compliance type as $C_i = \{W_i(0), W_i(1)\}$, we note that there are only 4 possible compliance types here:

	$W_i(1) = 0$	$W_i(1) = 1$
$W_i(0) = 0$	never taker	complier
$W_i(0) = 1$	defier	always taker

Our last assumption is that there are no defiers, i.e., $\mathbb{P}\left[C_i = \{1, 0\}\right] = 0$; this assumption is often also called monotonicity. In this case, one obtains a simple characterization of the IV estimand (10.2) by noting that

$$\begin{aligned}
&\mathbb{E}\left[Y_i \,\middle|\, Z_i = 1\right] - \mathbb{E}\left[Y_i \,\middle|\, Z_i = 0\right] \\
&\quad = \mathbb{E}\left[Y_i(W_i(1)) \,\middle|\, Z_i = 1\right] - \mathbb{E}\left[Y_i(W_i(0)) \,\middle|\, Z_i = 0\right] && \text{(exclusion)} \\
&\quad = \mathbb{E}\left[Y_i(W_i(1)) - Y_i(W_i(0))\right] && \text{(exogeneity)} \\
&\quad = \mathbb{E}\left[1\left(\{C_i = \text{complier}\}\right)\left(Y_i(1) - Y_i(0)\right)\right]. && \text{(no defiers)}
\end{aligned}$$

Thus, assuming that there actually exist some compliers (i.e., by relevance), we can apply Bayes' rule to conclude that

$$\begin{aligned}
\tau_{LATE} &= \frac{\mathbb{E}\left[Y_i \,\middle|\, Z_i = 1\right] - \mathbb{E}\left[Y_i \,\middle|\, Z_i = 0\right]}{\mathbb{E}\left[W_i \,\middle|\, Z_i = 1\right] - \mathbb{E}\left[W_i \,\middle|\, Z_i = 0\right]} \\
&= \mathbb{E}\left[Y_i(1) - Y_i(0) \,\middle|\, C_i = \text{complier}\right].
\end{aligned} \tag{10.3}$$

Although this is a very simple result, it already gives us some encouragement that IV methods can be interpreted in a non-parametric setting. The quantity identified in (10.3) is typically called the complier average treatment effect or, following Imbens and Angrist [1994], the local average treatment effect (LATE).

When the structural model (10.1) doesn't hold, the average treatment effect $\tau_{ATE} = \mathbb{E}\left[Y_i(1) - Y_i(0)\right]$ is clearly not identified without more data, because we don't have any observations on treated never takers, etc. However, under reasonable assumptions, IV methods let us estimate the most meaningful quantity we can identify here, namely the average treatment effect among those who are in fact "nudged" by the instrument.

Supply and demand Next, let's consider one of the classical settings for motivating instrumental variables regression: Estimating the effect of prices on demand. In many settings, it is of considerable interest to know the price

elasticity of demand, i.e., how demand would respond to price changes. In a typical marketplace, however, prices are not exogenous—rather, they arise from an interplay of supply and demand—and so estimating the elasticity requires an instrument.

One can formalize the relationship of supply and demand via potential outcomes as follows. For each marketplace $i = 1, ..., n$, there is a supply curve $S_i(p, z)$ and a demand curve $Q_i(p, z)$, corresponding to the supply (and respectively demand) that would arise given price $p \in \mathbb{R}$ and some instrument $z \in \{0, 1\}$ that may affect the marketplace (the instrument could, e.g., capture the presence of supply chain events that make production harder and thus reduce supply). For simplicity, we may take $S_i(\cdot, z)$ to be continuous and increasing and $Q_i(\cdot, z)$ to be continuous and decreasing.

Given this setting, suppose that first the instrument Z_i gets realized; then prices P_i arise by matching supply and demand, such that P_i is the unique solution to $S_i(P_i, Z_i) = Q_i(P_i, Z_i)$. The statistician observes the instrument Z_i, the market clearing price P_i ("the treatment") and the realized demand $Q_i = Q_i(P_i, Z_i)$ ("the outcome"). We say that Z_i is a valid instrument for measuring the effect of prices on demand if the following conditions hold:

- *Exclusion restriction.* The instrument only affects demand via supply, but not directly: $Q_i(p, z) = Q_i(p)$ for all p and z.
- *Exogeneity.* The instrument is randomized, $\{Q_i(p), S_i(p, z)\} \perp\!\!\!\perp Z_i$.
- *Relevance.* The instrument affects prices, $\text{Cov}[P_i, Z_i] \neq 0$.
- *Monotonicity.* $S_i(P_i, 1) \leq S_i(P_i, 0)$ almost surely.

Given this setting, we seek to estimate demand elasticity via (10.2).[2]

Now, although this may seem like a complicated setting, it turns out that the basic IV estimand admits a reasonably simple characterization. Suppose that $Q_i(p)$ is differentiable, and write $Q_i'(p)$ for its derivative.[3] Then,

$$\begin{aligned}\tau_{LATE} &= \frac{\mathbb{E}\left[Q_i \,\middle|\, Z_i = 1\right] - \mathbb{E}\left[Q_i \,\middle|\, Z_i = 0\right]}{\mathbb{E}\left[P_i \,\middle|\, Z_i = 1\right] - \mathbb{E}\left[P_i \,\middle|\, Z_i = 0\right]} \\ &= \frac{\int \mathbb{E}\left[Q_i'(p) \,\middle|\, P_i(0) \leq p \leq P_i(1)\right] \mathbb{P}\left[P_i(0) \leq p \leq P_i(1)\right] \, dp}{\int \mathbb{P}\left[P_i(0) \leq p \leq P_i(1)\right] \, dp},\end{aligned} \tag{10.4}$$

[2]To be precise, when studying demand elasticity we'd actually run this analysis with outcome $\log(Q_i)$ and treatment $\log(P_i)$. Here we'll ignore the logs for simplicity; introducing logs doesn't add any conceptual difficulties.

[3]The differentiability of $Q_i(\cdot)$ is not actually needed here: We've assumed that $Q_i(\cdot)$ is monotone increasing so that the distributional derivative must exist, and everything goes through with a distributional derivative.

i.e., the basic IV estimand can be written as a weighted average of the derivative of the demand function $Q_i(p)$ with respect to price p.

To verify this result, we first note that under the assumptions made here, i.e., that the instrument suppresses supply and that the supply and demand curves are monotone increasing and decreasing respectively, the instrument must have a monotone increasing effect on prices: $P_i(1) \geq P_i(0)$. Then,

$$\begin{aligned}
&\mathbb{E}\left[Q_i \,\middle|\, Z_i = 1\right] - \mathbb{E}\left[Q_i \,\middle|\, Z_i = 0\right] \\
&\quad = \mathbb{E}\left[Q_i(P_i(1)) \,\middle|\, Z_i = 1\right] - \mathbb{E}\left[Q_i(P_i(0)) \,\middle|\, Z_i = 0\right] && \text{(exclusion)} \\
&\quad = \mathbb{E}\left[Q_i(P_i(1)) - Q_i(P_i(0))\right] && \text{(exogen.)} \\
&\quad = \mathbb{E}\left[\int_{P_i(0)}^{P_i(1)} Q_i'(p)\ dp\right] && \text{(monot.)} \\
&\quad = \int \mathbb{E}\left[Q_i'(p) \,\middle|\, P_i(0) \leq p \leq P_i(1)\right] \mathbb{P}\left[P_i(0) \leq p \leq P_i(1)\right]\ dp, && \text{(Fubini)}
\end{aligned}$$

and the denominator in (10.4) can be characterized via similar means.

Threshold crossing and willingness to pay The other natural direction to extend our basic binary result with non-compliance is to the case of a real-valued instrument and a binary treatment. This setting could arise, for example, in a study of the effect of attending college ($W_i \in \{0, 1\}$) on lifetime income ($Y_i \in \mathbb{R}$), where we consider identification using an instrument Z_i that affects the cost of attending college (e.g., distance to the nearest college, or subsidies on tuition).

The standard way to model this setting is via a threshold crossing model: We assume that each subject has a latent and endogenous variable U_i such that

$$W_i = 1\left(\{U_i \geq c(Z_i)\}\right), \tag{10.5}$$

where $c(z)$ is some cutoff function depending on z. Concretely, in our example, one could interpret U_i as the i-th person's willingness to pay for college (which captures both their preferences and expected benefit anticipated from attending), while $c(z)$ represents the "cost" of attending as modulated by the instrument. Without loss of generality, we can take $U_i \sim \text{Unif}([0, 1])$, in which case $c(z) = 1 - \mathbb{P}\left[W_i = 1 \,\middle|\, Z_i = z\right]$. This boundary crossing structure yields a valid instrument under analogues to our usual assumptions:

- *Exclusion restriction.* There are potential outcomes $\{Y_i(0), Y_i(1)\}$ such that $Y_i = Y_i(W_i)$

- *Exogeneity.* The treatment assignment is randomized, meaning that $\{Y_i(0),\, Y_i(1),\, U_i\} \perp\!\!\!\perp Z_i$.
- *Relevance.* The threshold function $c(Z_i)$ has a non-trivial distribution.
- *Monotonicity.* The threshold function $c(z)$ is cadlag, non-decreasing.

Finally, define the marginal treatment effect

$$\tau(u) = \mathbb{E}\left[Y_i(1) - Y_i(0) \,\middle|\, U_i = u\right]. \tag{10.6}$$

Our goal is to show that IV methods recover a weighted average of the marginal treatment effect $\tau(u)$. Here, for convenience, we assume that the instrument is Gaussian, i.e., $Z_i \sim \mathcal{N}(0,\, 1)$. More general results without Gaussianity are given in Heckman and Vytlacil [2005].

Under these assumptions, one can check the following. Suppose that $\tau(u)$ is uniformly bounded, and that $\varphi(\cdot)$ is the standard Gaussian density. Then, the IV estimand (10.2) can be written as [4]

$$\tau_{LATE} = \frac{\sum_{z\in\mathcal{S}} \left(\int_{c_-(z)}^{c(z)} \tau(u)\ du\right) \varphi(z) + \int_{\mathbb{R}\setminus\mathcal{S}} \tau\left(c\left(z\right)\right) c'\left(z\right) \varphi\left(z\right)\ dz}{\sum_{z\in\mathcal{S}} \left(c(z) - c_-(z)\right) \varphi(z) + \int_{\mathbb{R}\setminus\mathcal{S}} c'\left(z\right) \varphi\left(z\right)\ dz}, \tag{10.7}$$

where $\mathcal{S} \subset \mathbb{R}$ is the set of discontinuity points of $c(\cdot)$ and $c_-(z) = \lim_{a\uparrow z} c(a)$. Thus, we immediately see that τ_{LATE} is a convex average of the marginal treatment function $\tau(u)$. We can get some further insight via examples:

Example: Single jump. Suppose that the threshold function $c(z)$ is constant with a single jump, i.e., $c(z) = c_0 + \delta_1 1\left(\{z \geq z_1\}\right)$. Then compliance types collapse into three principal strata: Never-takers with $U_i < c_0$, compliers with $c_0 \leq U_i < c_0 + \delta_1$, and always takers with $U_i \geq c_0 + \delta_1$. Furthermore, just as before, our estimand corresponds to the average treatment effect over the compliers (10.3).

Example: Multiple jumps. Now let there be K jumps, with cutoff function given by $c(z) = c_0 + \sum_{k=1}^{K} \delta_k 1\left(\{z \geq z_k\}\right)$. Then,

$$\tau_{LATE} = \frac{\sum_{k=1}^{K} \mathbb{E}\left[\tau(U_i) \,\middle|\, c_-(z_k) \leq U_i < c(z_k)\right] \left(c(z_k) - c_-(z_k)\right) \varphi(z_k)}{\sum_{k=1}^{K} \left(c(z_k) - c_-(z_k)\right) \varphi(z_k)}. \tag{10.8}$$

In other words, we recover a convex combination of average treatment effects over compliance strata defined by the jumps in $c(\cdot)$. These weights depend on

[4] Note that because $c(z)$ is monotone increasing it must also have bounded variation, and so we can write $c(z) = c_0 + \int_{-\infty}^{z} c'(a)\, da$ for some non-negative Lebesgue-measurable function $c'(z)$.

the size of the stratum (in U-space) and the density function of the instrument at z_k.

Example: Continuous cutoff. If the threshold function $c(z)$ has no jumps, then we recover the following weighted average of the marginal treatment effect function

$$\tau_{LATE} = \int_{\mathbb{R}} \tau\left(c\left(z\right)\right) c'\left(z\right) \varphi\left(z\right) \, dz \Big/ \int_{\mathbb{R}} c'\left(z\right) \varphi\left(z\right) \, dz. \tag{10.9}$$

In order to prove (10.7), the key task is in characterizing Cov $[Y_i,\, Z_i]$; an expression for the denominator of (10.2) can then be obtained via the same argument. First, note that

$$\begin{aligned}
\text{Cov}\left[Y_i,\, Z_i\right] &= \text{Cov}\left[Y_i(0) + (Y_i(1) - Y_i(0))W_i,\, Z_i\right] \\
&= \text{Cov}\left[(Y_i(1) - Y_i(0))W_i,\, Z_i\right] \\
&= \text{Cov}\left[(Y_i(1) - Y_i(0))1\left(\{U_i \geq c(Z_i)\}\right),\, Z_i\right] \\
&= \text{Cov}\left[\tau(U_i)1\left(\{U_i \geq c(Z_i)\}\right),\, Z_i\right],
\end{aligned}$$

where the first equality follows from the exclusion restriction, while the second and fourth follow from Assumption exogeneity. Now, write $H(z) = \mathbb{E}\left[\tau(U_i)1\left(\{U_i \geq c(z)\}\right)\right]$. Because Z_i is standard Gaussian, Lemma 1 of Stein [1981] implies that

$$\text{Cov}\left[H(Z_i),\, Z_i\right] = \mathbb{E}\left[H'(Z_i)\right], \tag{10.10}$$

where $H'(Z_i)$ denotes the distributional derivative of $H(\cdot)$. Furthermore, by Corollary 3.1 of Ambrosio and Dal Maso [1990],

$$-H'(z) = \begin{cases} \left(\int_{c_-(z)}^{c(z)} \tau(u) \, du\right) \delta_z & \text{for } z \in \mathcal{S}, \\ \tau\left(c\left(z\right)\right) c'\left(z\right) & \text{else}, \end{cases} \tag{10.11}$$

where δ_z is the Dirac delta-function at z. The representation (10.7) follows directly, noting that the minus-signs also appear in the denominator and thus get canceled out.

Estimating the marginal treatment effect Throughout this lecture, we've taken it as a given that we're going to target the estimand (10.2), and then have sought to interpret it in different settings. However, when we get to work with a continuous instrument, it's possible to target a wider variety of estimands. To this end, a first key result is that the marginal treatment effect

(10.6) is identified at continuity points of $c(z)$ via a "local IV" construction [Heckman and Vytlacil, 1999],

$$\tau(c(z)) = \frac{\frac{d}{dz}\,\mathbb{E}\left[Y_i \,\middle|\, Z_i = z\right]}{\frac{d}{dz}\,\mathbb{P}\left[W_i = 1 \,\middle|\, Z_i = z\right]}, \tag{10.12}$$

under regularity conditions whereby the ratio of derivatives is well defined. For intuition, note that this estimator has the same general form as the linear IV estimator (10.2), except that regression coefficients of Y_i and W_i on Z_i have been replaced with derivatives of the conditional response function. Then, once we have an identification result for the marginal treatment effect, we can use it to build estimators for various weighted averages of $\tau(u)$.

To verify (10.12), we start with the following observation: At any point u around which $c(Z_i)$ has continuous support,

$$\tau(c) = -\frac{d}{dc}\mathbb{E}\left[Y_i \,\middle|\, c(Z_i) = c\right]. \tag{10.13}$$

To check this fact, it suffices to note that

$$\begin{aligned}\mathbb{E}\left[Y_i \,\middle|\, c(Z_i) = c\right] &= \mathbb{E}\left[Y_i(0) + 1\left(\{U_i \geq c\}\right)\left(Y_i(1) - Y_i(0)\right) \,\middle|\, c(Z_i) = c\right] \\ &= \mathbb{E}\left[Y_i(0) + 1\left(\{U_i \geq c\}\right)\left(Y_i(1) - Y_i(0)\right)\right] = \mathbb{E}\left[Y_i(0)\right] + \int_c^1 \tau(u)du,\end{aligned}$$

where the first equality is due to (10.5) and the exclusion restriction, the second is due to exogeneity, and the third is an application of Fubini's theorem; (10.13) then follows via the fundamental theorem of calculus. Next, we can use the chain rule to check that

$$\frac{d}{dz}\mathbb{E}\left[Y_i \,\middle|\, Z_i = z\right] = \frac{d}{dc}\mathbb{E}\left[Y_i \,\middle|\, c(Z_i) = c\right] c'(z). \tag{10.14}$$

Finally, recall that by assumption $U_i \sim \text{Unif}([0,\, 1])$ independently of Z_i, and so $c'(z) = -(d/dz)\mathbb{P}\left[W_i = 1 \,\middle|\, Z_i = z\right]$. The result (10.12) follows by combining (10.13) with (10.14).

Bibliographic notes The idea of interpreting the results of instrumental variables analyses in terms of the local average treatment effect goes back to Imbens and Angrist [1994]. Our presentation of the analysis of clinical trials under non-compliance follows Angrist, Imbens, and Rubin [1996], while the local average treatment effect for supply-demand curves is discussed in Angrist, Graddy, and Imbens [2000].

Threshold crossing models of the form (10.5) have a long tradition in economics, where they are often discussed in the context of selection: People make choices if their (private) value from making that choice exceeds the cost. They are sometimes also called the Roy model following Roy [1951]. In the earlier literature, such selection models were often studied from a parametric point of view (without using instruments); for example, Heckman [1979] considers a problem where a treatment effect in a model of the type (10.5), and achieves identification by relying on joint normality of latent variable U_i and potential outcomes rather than on a source of exogenous randomness.

More recently, Heckman and Vytlacil [2005] have advocated for such selection models as a natural framework for understanding instrumental variables methods, and have studied methods that target a wide variety estimands beyond the LATE that may be more helpful in setting policy; in particular, the identification result (10.12) for the marginal treatment effect is discussed in Heckman and Vytlacil [1999]. For a discussion of semiparametrically efficient estimation of functions of the marginal treatment effect, see Kennedy, Lorch, and Small [2019].

Lecture 11
Policy Learning

So far, we've focused on methods for estimating causal effects in various statistical settings. In many application areas, however, the fundamental goal of performing a causal analysis isn't to estimate treatment effects, but rather to guide decision making: We want to understand treatment effects so that we can effectively prescribe treatment and allocate limited resources. The problem of learning optimal treatment assignment policies is closely related to—but subtly different from—the problem of estimating treatment heterogeneity. On one hand, policy learning appears easier: All we care about is assigning people to treatment or to control, and we don't care about accurately estimating treatment effects beyond that. On the other hand, when learning policies, we need to account for considerations that were not present when simply estimating treatment effects: Any policy we actually want to use must be simple enough we can actually deploy it, cannot discriminate on protected characteristics, should not rely on gameable features, etc. Today, we'll discuss how to learn treatment assignment policies by directly optimizing a relevant welfare criterion.

Policy learning For our purposes, a treatment assignment policy $\pi(x)$ is a mapping

$$\pi : \mathcal{X} \to \{0, 1\}, \tag{11.1}$$

such that individuals with features $X_i = x$ get treated if and only if $\pi(x) = 1$. Our goal is to find a policy that maximizes expected utility which, assuming potential outcomes $\{Y_i(0), Y_i(1)\}$ such that $Y_i = Y_i(W_i)$, can be written as (today, we'll always consider Y_i to be a utility to avoid discussions of risk preferences, etc.)

$$V(\pi) = \mathbb{E}\left[Y_i\left(\pi(X_i)\right)\right]. \tag{11.2}$$

Furthermore, today, we'll consider a setting where a subject-matter specialist has outlined a class Π of policies over which we're allowed to optimize.

Given this setting, for any class of policies Π, the optimal policy π^* (if it exists) is defined as

$$\pi^* = \operatorname{argmax} \{V(\pi') : \pi' \in \Pi\}, \tag{11.3}$$

while the regret of any other policy is

$$R(\pi) = \sup \{V(\pi') : \pi' \in \Pi\} - V(\pi). \tag{11.4}$$

Our goal is to learn a policy with guaranteed worst-case bounds of $R(\hat{\pi})$; this criterion is called the minimax regret criterion.

Exploring and exploiting To learn a good policy π, we need access to training data with exogenous assignment in the treatment assignment. For today, we'll assume we have access to $i = 1, \ldots, n$ IID samples $(X_i, Y_i, , W_i) \in \mathcal{X} \times \mathbb{R} \times \{0, 1\}$ sampled under unconfoundedness and overlap,

$$\begin{aligned} &\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp W_i \,\big|\, X_i, \qquad Y_i = Y_i(W_i), \\ &0 < \eta \leq e(X_i) \leq 1 - \eta < 1, \quad e(x) = \mathbb{P}\left[W_i = 1 \,\middle|\, X_i = x\right], \end{aligned} \tag{11.5}$$

and seek to use this data for learning a policy $\hat{\pi}$. Once we're done learning, we intend to deploy our policy: On our future samples we'll set $W_i = \hat{\pi}(X_i)$, and hope that the expected outcome $\mathbb{E}[Y_i]$ with $Y_i = Y_i(\hat{\pi}(X_i))$ will be large. In this second stage, there is no more randomness in treatment effects, so we cannot (non-parametrically) learn anything about causal effects anymore.

In engineering applications, the first phase is commonly caller "exploring" while the second phase is called "exploiting". There is a large literature on bandit algorithms that seek to merge the explore and exploit phases using a sequential algorithm; today, however, we'll focus on the "batch" case where the two phases are separate. Another major difference between our setting today and the bandit setting is that we've only assumed unconfoundedness (11.5), and in general will still need to worry about estimating propensity scores to eliminate confounding, etc. In contrast, in the bandit setting, exploration is carried out by the analyst, so the data collection process is more akin to a randomized trial.

Policy learning via empirical maximization If the optimal policy π^* is a maximizer of the true quality function $V(\pi)$ over $\pi \in \Pi$, then it is natural to learn $\hat{\pi}$ by maximizing an estimated quality function:

$$\hat{\pi} = \operatorname{argmax} \left\{ \widehat{V}(\pi) : \pi \in \Pi \right\}. \tag{11.6}$$

If we know the treatment propensities, then it turns out we have access to a simple, unbiased choice for $\widehat{V}(\pi)$ via inverse-propensity weighting:

$$\begin{aligned} \widehat{V}_{IPW}(\pi) &= \frac{1}{n}\sum_{i=1}^{n} \frac{1\left(\{W_i = \pi(X_i)\}\right) Y_i}{\mathbb{P}\left[W_i = \pi(X_i) \,\middle|\, X_i\right]}, \\ \hat{\pi}_{IPW} &= \operatorname{argmax}\left\{\widehat{V}_{IPW}\left(\pi\right) : \pi \in \Pi\right\}. \end{aligned} \tag{11.7}$$

In other words, we average outcome across those observations for which the sampled treatment W_i matches the policy prescription $\pi(X_i)$, and use inverse-propensity weighting to account for the fact that some relevant potential outcomes remain unobserved.

When the treatment propensities are known, we can readily check that, for any given policy π, the IPW estimate $\widehat{V}_{IPW}(\pi)$ is unbiased for $V(\pi)$:

$$\begin{aligned} \mathbb{E}\left[\widehat{V}(\pi)\right] &= \mathbb{E}\left[\frac{1\left(\{W_i = \pi(X_i)\}\right) Y_i}{\mathbb{P}\left[W_i = \pi(X_i) \,\middle|\, X_i\right]}\right] \\ &= \mathbb{E}\left[\frac{1\left(\{W_i = \pi(X_i)\}\right) Y_i(\pi(X_i))}{\mathbb{P}\left[W_i = \pi(X_i) \,\middle|\, X_i\right]}\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\frac{1\left(\{W_i = \pi(X_i)\}\right)}{\mathbb{P}\left[W_i = \pi(X_i) \,\middle|\, X_i\right]} \,\middle|\, X_i\right] \mathbb{E}\left[Y_i(\pi(X_i)) \,\middle|\, X_i\right]\right] \\ &= \mathbb{E}\left[Y_i(\pi(X_i))\right] = V(\pi), \end{aligned} \tag{11.8}$$

where the second equality follows by consistency of potential outcomes and the third by unconfoundedness.

Policy learning as weighted classification The above unbiasedness result suggests that $\widehat{V}_{IPW}(\pi)$ may be a reasonable estimate of the policy value. However, our approach to learning via (11.7) doesn't just involve evaluation a single policy π; rather, we learn by taking an argmax. Thus, before using the estimator $\hat{\pi}_{IPW}$, it's important to understand the properties of this maximization step—both statistically and computationally.

To this end, it's helpful to reparametrize our problem, starting from the value function itself. The value function can be decomposed as $V(\pi) = \mathbb{E}\left[Y_i(0)\right] + \mathbb{E}\left[\left(Y_i(1) - Y_i(0)\right)\pi(X_i)\right]$, highlighting its dependence on both the baseline effect and the average treatment effect among those treated by $\pi(\cdot)$. Now, the baseline effect is unaffected by policy choice, and so it's helpful to re-center our objective such as to focus on the part of the problem we can work

with, namely the conditional average treatment effect:

$$\begin{aligned} A(\pi) &= 2\mathbb{E}\left[Y_i\left(\pi(X_i)\right)\right] - \mathbb{E}\left[Y_i(0) + Y_i(1)\right], \\ &= \mathbb{E}\left[\left(2\pi(X_i) - 1\right)\tau(X_i)\right] \end{aligned} \tag{11.9}$$

Here, A stands for the "advantage" of the policy $\pi(\cdot)$. Of course, π^* is still the maximizer of $A(\pi)$ over $\pi \in \Pi$, etc. We can similarly re-express the IPW objective: $\hat{\pi}_{IPW}$ maximizes $\widehat{A}_{IPW}(\pi)$, where

$$\begin{aligned} \widehat{A}_{IPW}(\pi) &= 2\,\widehat{V}_{IPW}(\pi) - \frac{1}{n}\sum_{i=1}^{n}\left(\frac{W_i Y_i}{e(X_i)} + \frac{(1-W_i)Y_i}{1-e(X_i)}\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(2\pi(X_i) - 1\right)\left(\frac{W_i Y_i}{e(X_i)} - \frac{(1-W_i)Y_i}{1-e(X_i)}\right), \end{aligned} \tag{11.10}$$

where by an analogous derivation to (11.8) we see that $\widehat{A}_{IPW}(\pi)$ is unbiased for $A_{IPW}(\pi)$.

The new for (11.10) gives us several insights on the for of the IPW objective for policy learning. First, for intuition, we note that

$$\widehat{A}_{IPW}(\pi) = \frac{1}{n}\sum_{i=1}^{n}\left(2\pi(X_i) - 1\right)\Gamma_i^{IPW}, \quad \Gamma_i^{IPW} = \frac{W_i Y_i}{e(X_i)} - \frac{(1-W_i)Y_i}{1-e(X_i)}, \tag{11.11}$$

where the Γ_i^{IPW} are IPW-scores familiar from our analysis of average treatment effect estimation; specifically, the IPW estimate of the average treatment effect is $\hat{\tau}_{IPW} = n^{-1}\sum_{i=1}^{n}\Gamma_i^{IPW}$. Thus, we see that $\widehat{A}_{IPW}(\pi)$ is like an IPW estimator for the ATE, except we "earn" the treatment effect for the i-th sample when $\pi(X_i) = 1$, and "pay" the treatment effect when $\pi(X_i) = 0$.

Meanwhile, for the purpose of optimization, we can write the objective as

$$\widehat{A}_{IPW}(\pi) = \frac{1}{n}\sum_{i=1}^{n}\underbrace{\left(2\pi(X_i) - 1\right)\operatorname{sign}\left(\Gamma_i\right)}_{\text{classification objective}}\underbrace{\left|\Gamma_i\right|}_{\text{sample weight}}. \tag{11.12}$$

In other words, maximizing $\widehat{A}_{IPW}(\pi)$ is equivalent to optimizing a weighted classification objective. This means that we can use any software for weighted minimization of a classification loss to learn $\hat{\pi}_{IPW}$.[1]

[1]As a note of caution: We've found that policy learning via empirical maximization is computationally equivalent to weighted optimization of a classification objective. In practice, however, we often carry out classification by optimization a surrogate objective (rather than the basic classification objective), e.g., the using the hinge or logistic loss, and so it may be tempting to seek to learn policies by weighted minimization of a similar surrogate loss. The guarantees presented here, however, do not extend to such an approach. For example, it's possible to design situations where learning with a "logistic" surrogate for (11.12) makes us prioritize people who would benefit the least from treatment (rather than the most); see Wager [2020b] for a discussion.

Furthermore, from this connection, we directly obtain regret bounds for the learned policy. If we assume that $|Y_i| \leq M$ and $\eta \leq e(X_i) \leq 1 - \eta$, such as to make the weights Γ_i bounded, and assume that Π has a bounded Vapnik-Chervonenkis dimension, then the regret of $\hat{\pi}$ is bounded as

$$R\left(\hat{\pi}_{IPW}\right) = \mathcal{O}_P\left(\frac{M}{\eta}\sqrt{\frac{\text{VC}(\Pi)}{n}}\right),\ \hat{\pi}_{IPW} = \text{argmax}_{\pi\in\Pi}\left\{\widehat{A}_{IPW}\left(\pi\right)\right\}, \quad (11.13)$$

where $\text{VC}(\Pi)$ denotes the VC-dimension of Π.

Efficient scoring rules for policy learning Although the IPW policy learning method discussed above has some nice properties (e.g., $\sqrt{n}$ regret consistency), we may still ask whether it is the best possible such method. To get a better understanding of this issue, it is helpful to turn back to our discussions of ATE estimation.

In order to learn a good policy $\hat{\pi}$, it is intuitively helpful to start with a good method $\widehat{A}(\pi)$ for evaluating the quality of individual policies π. And here, we can start by noting that

$$\begin{aligned} A(\pi) &= 2\mathbb{E}\left[Y_i\left(\pi(X_i)\right)\right] - \mathbb{E}\left[Y_i(0) + Y_i(1)\right] \\ &= \mathbb{E}\left[Y_i\left(\pi(X_i)\right)\right] - \mathbb{E}\left[Y_i\left(1 - \pi(X_i)\right)\right]. \end{aligned} \quad (11.14)$$

In other words, $A(\pi)$ is the ATE in an experiment where we compare deploying the policy $\pi(\cdot)$ to and experiment where we always deploy the *opposite* of $\pi(\cdot)$.

Now, given this formulation as an ATE estimation problem, we know that the oracle IPW estimator is OK, but not efficient. The oracle AIPW estimator $\widehat{A}^*_{AIPW}(\pi)$ that estimates $A(\pi)$ by averaging an efficient score attains the semiparametric efficiency bound; and, in our case,

$$\begin{aligned} \widehat{A}^*_{AIPW}(\pi) &= \frac{1}{n}\sum_{i=1}^{n}\left(2\pi(X_i) - 1\right)\Gamma^*_i, \\ \Gamma^*_i &:= \mu_{(1)}(X_i) - \mu_{(0)}(X_i) + W_i\frac{Y_i - \mu_{(1)}(X_i)}{e(X_i)} \\ &\qquad - \left(1 - W_i\right)\frac{Y_i - \mu_{(0)}(X_i)}{1 - e(X_i)}. \end{aligned} \quad (11.15)$$

Furthermore, assuming the existence of $o_P(n^{-1/4})$-consistent regression adjustments for $\hat{\mu}_{(w)}(x)$ and $\hat{e}(x)$ we can construct a doubly robust estimator that

emulates the efficient oracle:

$$\begin{aligned}
\widehat{A}_{AIPW}(\pi) &= \frac{1}{n}\sum_{i=1}^{n}(2\pi(X_i)-1)\,\widehat{\Gamma}_i, \\
\widehat{\Gamma}_i &:= \hat{\mu}_{(1)}^{(-k(i))}(X_i) - \hat{\mu}_{(0)}^{(-k(i))}(X_i) \\
&\quad + W_i\frac{Y_i - \hat{\mu}_{(1)}^{(-k(i))}(X_i)}{\hat{e}^{(-k(i))}(X_i)} - (1-W_i)\frac{Y_i - \hat{\mu}_{(0)}^{(-k(i))}(X_i)}{1-\hat{e}^{(-k(i))}(X_i)},
\end{aligned} \tag{11.16}$$

and note that this is also a weighted classification objective.

We already know from lecture 3 that $\widehat{A}_{AIPW}(\pi)$ is pointwise asymptotically equivalent to $\widehat{A}^*_{AIPW}(\pi)$, i.e., for any fixed policy π the difference between the two quantities decays faster than $1/\sqrt{n}$. However, more is true: If Π is a VC class, then

$$\sqrt{n}\sup\left\{\left|\widehat{A}_{AIPW}(\pi) - \widehat{A}^*_{AIPW}(\pi)\right| : \pi\in\Pi\right\} \to_p 0. \tag{11.17}$$

This result, along with an empirical process concentration argument then imply that the regret of policy learning with the AIPW-scoring rule is bounded on the order of

$$\begin{aligned}
&R\left(\hat{\pi}_{AIPW}\right) = \mathcal{O}_P\left(\sqrt{\frac{V^*\,\mathrm{VC}(\Pi)}{n}}\right), \quad \hat{\pi}_{AIPW} = \mathrm{argmax}_{\pi\in\Pi}\left\{\widehat{A}_{AIPW}\left(\pi\right)\right\}, \\
&V^* = \mathbb{E}\left[\tau^2(X_i)\right] + \mathbb{E}\left[\frac{\mathrm{Var}\left[Y_i(0)\,\middle|\,X_i\right]}{1-e(X_i)}\right] + \mathbb{E}\left[\frac{\mathrm{Var}\left[Y_i(1)\,\middle|\,X_i\right]}{e(X_i)}\right].
\end{aligned}$$

See Athey and Wager [2017] for details, as well as lower bounds. Effectively, the above bound is optimal in a regime where treatment effects just barely peak out of the noise.

The role of the policy class Π This problem setup may appear unusual. We started with a non-parametric model (i.e., $\mu_{(w)}(x)$ and $e(x)$ can be generic), in which case the Bayes-optimal treatment assignment rule is simply $\pi_{bayes}(x) = 1\left(\{\tau(x) > 0\}\right)$. However, from this point, our goal was not to find a way to approximate $\pi_{bayes}(x)$; rather, given another, pre-specified class of policies Π, we want to learn a nearly regret-optimal representative from Π. For example, Π could consist of linear decision rules, k-sparse decision rules, depth-ℓ decision trees, etc. Note, in particular, that we never assumed that $\pi_{bayes}(\cdot)\in\Pi$.

The reason for this tension is that the features X_i play two distinct roles here. First, the X_i may be needed to achieve unconfoundedness

$$\{Y_i(0),\, Y_i(1)\} \perp\!\!\!\perp W_i \,\big|\, X_i. \tag{11.18}$$

In general, the more pre-treatment variables we have access to, the more plausible unconfoundedness becomes. In order to have a credible model of nature, it's good to have flexible, non-parametric models for $e(x)$ and $\mu_{(w)}(x)$ using a wide variety of features.

On the other hand, when we want to deploy a policy $\pi(\cdot)$, we should be much more careful about what features we use to make decisions and the form of the policy $\pi(\cdot)$:

- We should not use certain features, e.g., features that are difficult to measure in a deployed system, features that are gameable by participants in the system, or features that correspond to legally protected classes.
- We may have budget constraints (e.g., at most 15% of people get treated), or marginal budget constraints (e.g., the total amount of funds allocated to each state stays fixed, but we may re-prioritize funds within states).
- We may have functional form constraints on $\pi(\cdot)$ (e.g., if the policy needs to be communicated to employees in a non-electronic format, or audited using non-quantitative methods).

Given any such constraints set by a practitioner, we can construct a class of allowable policies Π that respects these feature exclusion, budget, and functional form constraints.

Bibliographic notes The idea behind our discussion today was that, when learning policies, the natural quantity to focus on is regret as opposed to, e.g., squared-error loss on the conditional average treatment effect function. This point is argued for in Manski [2004]. For a discussion of exact minimax regret policy learning with discrete covariates, see Stoye [2009].

The insight that policy learning under unconfoundedness can be framed as a weighted classification problem—and that we can adapt well known result results from empirical risk minimization to to derive useful regret bounds—appears to have been independently discovered in statistics [Zhao, Zeng, Rush, and Kosorok, 2012], computer science [Swaminathan and Joachims, 2015], and economics [Kitagawa and Tetenov, 2018]. Properties of policy learning with doubly robust scoring rules are derived in Athey and Wager [2017]. The latter paper also considers policy learning in more general settings, such as with "nudge" interventions to continuous treatments or with instruments used to identify the effects of endogenous treatments.

Today, we've discussed rates of convergence that scale as $\sqrt{\text{VC}(\Pi)/n}$. This is the optimal rate of convergence we can get if seek guarantees that are uniform over $\tau(x)$; and the rates are sharp when the strength of the treatment

effects decays with sample size at rate $1/\sqrt{n}$. However, if we consider asymptotics for fixed choices of $\tau(x)$, then super-efficiency phenomena appear and we can obtain faster than $1/\sqrt{n}$ rates [Luedtke and Chambaz, 2017]; this phenomenon is closely related to "large margin" improvements to regret bounds for classification via empirical risk minimization.

Finally, the topic of policy learning is an active area with many recent advances. Bertsimas and Kallus [2020] extend the principle of learning policies by optimizing a problem-specific empirical value function to a wide variety of settings, e.g., inventory management. Luedtke and van der Laan [2016] discuss inference for the value of the optimal policy. Finally, Kallus and Zhou [2018] consider the problem of learning policies in a way that is robust to potential failures of unconfoundedness.

Lecture 12
Evaluating Dynamic Policies

In many real-world applications, "treatment" is just a one-shot decision that can be set to 0 or 1, but rather an ongoing set of decisions. Consider, for example, the case of antiretroviral therapy (ART) for HIV-positive patients. It is understood that HIV reduces CD4 white blood cell count, and that patients are at risk of contracting AIDS-defined illnesses once CD4 count is low; the use of ART can help preserve CD4 counts, but it is a very intensive form of medication. Traditional guidelines for treating HIV recommend beginning ART when CD4 count is low; but recent guidelines recommend ART as soon as HIV is diagnosed. To study problems like this, we need to allow for treatment assignment policies that vary across time, and respond to time-varying covariates (e.g., CD4 count).

The goal of today's lecture is to provide a brief introduction to working with dynamic treatment policies in the context of the potential outcomes model. Unlike in our earlier discussion of panel data, we'll allow for generic dynamics (e.g., a poor treatment choice yesterday may worsen a patient's outcomes today, which in turn will make more likely the adoption of an aggressive treatment regime tomorrow). The problem of evaluating and learning dynamic policies is often called reinforcement learning in the engineering community.

Statistical setting As always, our statistical analysis starts with the specification of potential outcomes, a target estimand, and an identifying assumption. Suppose we have data on $i = 1, \ldots, n$ IID patients, observed at times $t = 1, \ldots, T$. At each time point, we observe a set of (time-varying) covariates X_{it} as well as a treatment assignment $W_{it} \in \{0, 1\}$. Finally, once we reach time T, we also observe an outcome $Y_i \in \mathbb{R}$.

To reflect the dynamic structure of the problem, we let any time-varying observation depend on all past treatment assignments. Thus, for each $X_{it} \in \mathcal{X}_t$, we define 2^{t-1} potential outcomes $X_{it}(w_{1:(t-1)})$ such that $X_{it} = X_{it}(W_{i(1:(t-1))})$, while for the final outcome we have 2^T potential outcomes $Y_i(w_{1:T})$ such that

$Y_i = X_{it}(W_{i(1:T)})$. Finally, the the treatment assignment W_{it} may depend on $X_{i(1:t)}$ as well as past values of treatment; and, to reflect this possibility, we need to define potential outcomes for treatment, $W_{it}(w_{1:(t-1)})$, such that $W_{it} = W_{it}(W_{i(1:(t-1))})$.

Next, we need to define an estimand. In the dynamic setting, the number of potential treatment allocation rules grows exponentially with the horizon T, and so does the number of questions we can ask. Some common estimands are:

- Evaluate a fixed treatment choice, i.e., for some pre-specified $w \in \{0, 1\}^T$, estimate
$$V(w) = \mathbb{E}\left[Y_i(w)\right]. \tag{12.1}$$
- Evaluate a treatment policy. For this purpose, a policy is a set of mappings $\pi_t : \mathcal{X}_t \to \{0, 1\}$ that, at each time point sets treatment $W_{it} = \pi(X_{it})$. Then, the value of the policy π is
$$V(\pi) = \mathbb{E}\left[Y_i(\pi_1(X_{i1}),\, \pi_2(X_{i1},\, \pi_1(X_{i1}),\, X_{i2}(\pi_1(X_{i1})),\, \ldots)\right]. \tag{12.2}$$
This notation is fairly verbose, because it allows time-t covariates (which inter into our choice of time-t action) to depend on past treatments.

There are also several questions that can be raised in terms of randomized treatment assignment rules (including perturbations to the treatment assignment distribution used to collect the data).

There are several natural unconfoundedness-type assumptions that can be used to identify our target estimands. One option is to posit sequential unconfoundedness (or sequential ignorability),

$$\{(\text{potential outcomes after time } t)\} \perp\!\!\!\perp W_{it} \,\big|\, \{(\text{History at time } t)\}, \tag{12.3}$$

i.e., we assume that W_{it} is always "uncounfounded" in the usual sense given data that was collected up to time t. A stronger assumption is to posit complete randomization

$$\{(\text{all potential outcomes})\} \perp\!\!\!\perp W_{1:T}. \tag{12.4}$$

Complete randomization leads to easier statistical analysis, but may force us to explore some unreasonable treatment assignment rules (e.g., what if you enroll someone in a cancer trial, and they're randomized to the arm "start chemotherapy in one year", but after one year it turns out they're already cured and so don't need chemotherapy). Today, we'll focus on methods that work under sequential unconfoundedness.

Treatment-confounder feedback Working with sequential unconfoundedness gives rise to a subtle difficulty that is not present in the basic (single-period) setting, namely treatment-confounder feedback.

To see what may go wrong, consider the following simple example adapted from Hernán and Robins [2020], modeled after an ART trial with $T = 2$ time periods. Here, $X_{it} \in \{0, 1\}$ denotes CD4 count (1 is low, i.e., bad), and suppose that $X_{i1} = 0$ for everyone (no one enters the trial very sick), and X_{i1} is randomized with probability 0.5 of receiving treatment. Then, at time period 2, we observe X_{i2} and assign treatment $X_{i2} = 1$ with probability 0.4 if $X_{i2} = 0$ and with probability 0.8 if $X_{i2} = 1$. In the end, we collect a health outcome Y. This is a sequential randomized experiment.

n	X_{i1}	W_{i1}	X_{i2}	W_{i2}	Mean Y
2400	0	0	0	0	84
1600	0	0	0	1	84
2400	0	0	1	0	52
9600	0	0	1	1	52
4800	0	1	0	0	76
3200	0	1	0	1	76
1600	0	1	1	0	44
6400	0	1	1	1	44

We observe data as in the Table above (the last column is the mean outcome for everyone in that row). Our goal is to estimate $\tau = \mathbb{E}\left[Y\left(\underline{1}\right) - Y\left(\underline{0}\right)\right]$, i.e., the difference between the always treat and never treat rules. How should we do this? As a preliminary, it's helpful to note that the treatment obviously does nothing. In the first time period,

$$\mathbb{E}\left[Y_i \,\middle|\, W_{i1} = 0\right] = \mathbb{E}\left[Y_i \,\middle|\, W_{i1} = 1\right] = 60,$$

and this is obviously a causal quantity (since W_{i1} was randomized). Moreover, in the second time period we see by inspection that

$$\mathbb{E}\left[Y_i \,\middle|\, W_{i2} = 0,\, W_{i1} = w_1,\, X_{i2} = x\right] = \mathbb{E}\left[Y_i \,\middle|\, W_{i2} = 1,\, W_{i1} = w_1,\, X_{i2} = x\right],$$

for all values of w_1 and x, and again the treatment does nothing.

However, some simple estimation strategies that served us well in the non-dynamic setting do not get the right answer. In particular, here are some strategies that do not get the right answer:

- Ignore adaptive sampling, and use

$$\begin{aligned}\hat{\tau} &= \widehat{\mathbb{E}}\left[Y \,\middle|\, W = \underline{1}\right] - \widehat{\mathbb{E}}\left[Y \,\middle|\, W = \underline{0}\right] \\ &= \frac{6400 \times 44 + 3200 \times 76}{6400 + 3200} - \frac{2400 \times 52 + 2400 \times 84}{2400 + 2400} \\ &= 54.7 - 68 = -13.3.\end{aligned}$$

- Stratify by CD4 count at time 2, to control for adaptive sampling:

$$\begin{aligned}\hat{\tau}_0 &= \mathbb{E}\left[Y \,\middle|\, W = \underline{1},\, X_{i2} = 0\right] - \mathbb{E}\left[Y \,\middle|\, W = \underline{0},\, X_{i2} = 0\right] = 76 - 84 = -8 \\ \hat{\tau}_1 &= \mathbb{E}\left[Y \,\middle|\, W = \underline{1},\, X_{i2} = 1\right] - \mathbb{E}\left[Y \,\middle|\, W = \underline{0},\, X_{i2} = 1\right] = 44 - 52 = -8 \\ \hat{\tau} &= \frac{(3200 + 2400)\hat{\tau}_0 + (6400 + 2400)\hat{\tau}_1}{3200 + 2400 + 6400 + 2400} = -8.\end{aligned}$$

The problem with the first strategy is obvious (we need to correct for biased sampling). But the problem with the second strategy is more subtle. We know via sequantial randomization that

$$Y_i(\cdots) \perp\!\!\!\perp W_{i2} \,\big|\, X_{i2},$$

and this seems to justify stratification. But what we'd actually need for stratification is:

$$Y_i(\cdots) \perp\!\!\!\perp (W_{i1},\, W_{i2}) \,\big|\, X_{i2},$$

and this is *not* true by design. To see what could go wrong, imagine that there are 3 types of people (stable, responder, acute), and tabulate their time-2 CD4 values as follows (these categories are usually called principal strata):

	$W_{i1} = 0$	$W_{i1} = 1$
stable	$X_{i2} = 0$	$X_{i2} = 0$
responder	$X_{i2} = 1$	$X_{i2} = 0$
acute	$X_{i2} = 1$	$X_{i2} = 1$

These principal strata are unobservable (just like compliance types in IV analyses), but can still provide insights. For example:

- $\mathbb{E}\left[Y \,\middle|\, W = \underline{1},\, X_{i2} = 0\right]$ is an average over stable or responder patients, whereas $\mathbb{E}\left[Y \,\middle|\, W = \underline{0},\, X_{i2} = 0\right]$ is simply an average over stable patients. So the difference $\hat{\tau}_0$ is not estimating a proper causal quantity.

- $\mathbb{E}\left[Y \,\middle|\, W = \underline{1},\, X_{i2} = 1\right]$ is an average over acute patients, whereas in contrast $\mathbb{E}\left[Y \,\middle|\, W = \underline{0},\, X_{i2} = 1\right]$ is an average over responder or acute patients. So the difference $\hat{\tau}_1$ is not estimating a proper causal quantity.

In other words, in sequentially randomized trials, stratification does not control for confounding.

Sequential inference for sequential ignorability Since stratification doesn't work, we now move to study a family of approaches that do. Here, we focus on estimating the value of a policy $V(\pi)$ as in (12.2); note that evaluating a fixed treatment sequence is a special case of this strategy. To this end, it's helpful to define some more notation:

- We denote by $\mathcal{F}_t = \sigma\left(\{X_1, W_1, \ldots, W_{t-1}, X_t\}\right)$ the filtration containing all information until the period-t treatment is chosen.
- We use the shorthand $\mathbb{E}_\pi$ to denote expectations with treatment set using policy π such that, e.g., (12.2) becomes $V(\pi) = \mathbb{E}_\pi[Y]$.
- We define the value function

$$V_{\pi,t}(X_1, W_1, \ldots, W_{t-1}, X_t) = \mathbb{E}_\pi\left[Y \,\middle|\, \mathcal{F}_t\right] \tag{12.5}$$

  that measures the expected reward we'd get if we were to start following π given our current state as captured by $\mathcal{F}_t$.

This notation lets us concisely express a helpful principle behind fruitful estimation of $V(\pi)$ (note that, given (12.5), we could say that the overall value of a policy is $V_{\pi,0}$): By the chain rule, we see that

$$\begin{aligned}\mathbb{E}_\pi\left[V_{\pi,t+1}(X_1, W_1, \ldots, W_t, X_{t+1}) \,\middle|\, \mathcal{F}_t\right] &= \mathbb{E}_\pi\left[\mathbb{E}_\pi\left[Y \,\middle|\, \mathcal{F}_{t+1}\right] \,\middle|\, \mathcal{F}_t\right] \\ &= \mathbb{E}_\pi\left[Y \,\middle|\, \mathcal{F}_t\right] = V_{\pi,t}(X_1, W_1, \ldots, W_{t-1}, X_t).\end{aligned} \tag{12.6}$$

The implication is that, given a good estimate of $V_{\pi,t+1}$, all we need to be able to do is to get a good estimate of $V_{\pi,t}$; then we can recurse our way backwards to $V(\pi)$. The question is then how we choose to act on this insight.

Finally, the $\mathbb{E}_\pi$ notation from (12.5) lets us also capture sequential ignorability in terms of more tractable notation. We can always factor the joint distribution of $(X_1, W_1, \ldots, X_T, W_T, X_{T+1})$ as (where we used the short-hand $Y = X_{T+1}$)

$$\begin{aligned}&\mathbb{P}_\pi[X_1, W_1, \ldots, X_T, W_T, Y] \\ &\quad = \mathbb{P}_\pi[X_1] \prod_{t=1}^{T} \mathbb{P}_\pi\left[W_t \,\middle|\, \mathcal{F}_t\right] \mathbb{P}_\pi\left[X_{t+1} \,\middle|\, \mathcal{F}_t, W_t\right].\end{aligned} \tag{12.7}$$

Here, unconfoundedness implies that terms in the factorization that don't integrate over W_t don't depend on the policy π, i.e.,

$$\begin{aligned}\mathbb{P}_\pi[X_1] &= \mathbb{P}[X_1] \\ \mathbb{P}_\pi\left[X_{t+1} \,\middle|\, \mathcal{F}_t, W_t\right] &= \mathbb{P}\left[X_{t+1} \,\middle|\, \mathcal{F}_t, W_t\right],\end{aligned} \tag{12.8}$$

for all policies π.

Inverse-propensity weighting A first step in making (12.6) useful is taking a change of measure. Given our training sample, it's easy to measure expectations $\mathbb{E}$ according to the training treatment assignment distribution; but here, we instead seek expectations with respect to the "off-policy" distribution, with treatment assigned according to π. To carry out the change of measure, we note that (recall that X_1, W_1, ..., W_{t-1}, X_t are fixed by the conditioning event)

$$\begin{aligned}
&V_{\pi,t}(X_1, W_1, \ldots, W_{t-1}, X_t) \\
&\quad = \mathbb{E}_\pi \left[V_{\pi,t+1}(X_1, W_1, \ldots, W_t, X_{t+1}) \,\middle|\, \mathcal{F}_t \right] \\
&\quad = \mathbb{E} \left[\frac{\mathbb{P}_\pi \left[W_t, X_{t+1} \,\middle|\, \mathcal{F}_t \right]}{\mathbb{P} \left[W_t, X_{t+1} \,\middle|\, \mathcal{F}_t \right]} V_{\pi,t+1}(X_1, W_1, \ldots, W_t, X_{t+1}) \,\middle|\, \mathcal{F}_t \right] \\
&\quad = \mathbb{E} \left[\frac{\mathbb{P}_\pi \left[W_t \,\middle|\, \mathcal{F}_t \right] \mathbb{P}_\pi \left[X_t \,\middle|\, \mathcal{F}_t, W_t \right]}{\mathbb{P} \left[W_t \,\middle|\, \mathcal{F}_t \right] \mathbb{P} \left[X_t \,\middle|\, \mathcal{F}_t, W_t \right]} V_{\pi,t+1}(X_1, W_1, \ldots, W_t, X_{t+1}) \,\middle|\, \mathcal{F}_t \right] \\
&\quad = \mathbb{E} \left[\frac{1\left(\{W_t = \pi_t(\ldots, X_t)\}\right)}{\mathbb{P} \left[W_t = \pi_t(\ldots, X_t) \,\middle|\, \mathcal{F}_t \right]} V_{\pi,t+1}(X_1, W_1, \ldots, W_t, X_{t+1}) \,\middle|\, \mathcal{F}_t \right],
\end{aligned}$$

where the key step here was the last equality which used the fact that, by unconfoundedness, $\mathbb{P}_\pi \left[X_t \,\middle|\, \mathcal{F}_t, W_t \right] = \mathbb{P} \left[X_t \,\middle|\, \mathcal{F}_t, W_t \right]$.

Now, to turn this fact into an estimator of $V(\pi)$, we write down our change of measure relationship for each $t = 1$, ..., T:

$$\begin{aligned}
V(\pi) &= \mathbb{E} \left[V_{\pi,1}(X_1) \right], \\
V_{\pi,1}(X_1) &= \mathbb{E} \left[\frac{1\left(\{W_1 = \pi_1(X_1)\}\right)}{\mathbb{P} \left[W_1 = \pi_1(X_1,) \right]} V_{\pi,2}(X_1, W_2, X_2) \,\middle|\, \mathcal{F}_1 \right],
\end{aligned}$$

etc. Then we can start backwards-substituting, always replacing expressions in terms of V_t for ones in terms of V_{t+1}, until only $V_{\pi,\,T+1}(\cdots) = \mathbb{E}_\pi \left[Y \,\middle|\, \mathcal{F}_{T+1} \right] = Y$ is left. Finally, we recover

$$V(\pi) = \mathbb{E} \left[\prod_{t=1}^{T} \frac{1\left(\{W_t = \pi_t(\ldots, X_t)\}\right)}{\mathbb{P} \left[W_t = \pi_t(\ldots, X_t) \,\middle|\, \mathcal{F}_t \right]} Y \right], \tag{12.9}$$

which leads naturally leads to an IPW-type estimator

$$\begin{aligned}
\widehat{V}_{IPW}(\pi) &= \frac{1}{n} \sum_{i=1}^{n} \gamma_{iT}(\pi) Y_i, \\
\gamma_{it}(\pi) &= \gamma_{i(t-1)}(\pi) \, \frac{1\left(\{W_t = \pi_t(\ldots, X_t)\}\right)}{\mathbb{P} \left[W_t = \pi_t(\ldots, X_t) \,\middle|\, \mathcal{F}_t \right]},
\end{aligned} \tag{12.10}$$

where $\gamma_{i0}(\pi) = 0$. This estimator averages outcomes whose treatment trajectory exactly matches π, while applying an IPW correction for selection effects due to measured (time-varying) confounders. Our derivation immediately implies that the IPW estimator is unbiased if we know the inverse-propensity weights γ_{iT} exactly.

Backwards regression adjustment As always, the other way to leverage (12.6) and sequential unconfoundedness is via a regression adjustment. This approach again proceeds by backwards iteration. First, for $t = T$, we can use sequential unconfoundedness to check that

$$\begin{aligned} V_{\pi,T}(X_1, W_1, \ldots, X_T) &= \mathbb{E}_\pi \left[Y \,\middle|\, \mathcal{F}_T \right] \\ &= \mathbb{E}_\pi \left[Y \,\middle|\, \mathcal{F}_T, W_T = \pi_T(X_1, W_1, \ldots, X_T) \right] \\ &= \mathbb{E} \left[Y \,\middle|\, \mathcal{F}_T, W_T = \pi_T(X_1, W_1, \ldots, X_T) \right]. \end{aligned} \tag{12.11}$$

We can then take this as a non-parametric regression problem, and seek to learn $\widehat{V}_{\pi,T}(X_1, W_1, \ldots, X_T)$. Then, in the recursive step, we note that if we have a reasonable estimate of $\widehat{V}_{\pi,\,t+1}$, then

$$V_{\pi,t}(X_1, W_1, \ldots, X_t) \approx \mathbb{E} \left[\widehat{V}_{\pi,\,t+1}(\ldots, X_{t+1}) \,\middle|\, \mathcal{F}_t, W_t = \pi_t(\ldots, X_t) \right]. \tag{12.12}$$

We can again keep recursing backwards, until we recover an estimate of $\widehat{V}(\pi)$. Unlike IPW, formal analysis of the regression adjustment method is more delicate, as we need to carefully quantify how regression errors propagate as we iterate backwards. Note that, in the reinforcement learning literature, the backwards-recursive regression based approach is typically referred to as Q-learning.

A doubly robust estimator Where there's an IPW and a regression based estimator, there's going to be a doubly robust estimator also. To construct one, it's helpful to consider the last step of the regression-estimator (12.12): We've derived a good value estimate $\widehat{V}_{\pi,\,1}(X_1)$, and conclude by setting

$$\widehat{V}_{REG}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \widehat{V}_{\pi,\,1}(X_{i1}). \tag{12.13}$$

Now, what would a one-step doubly robust correction look like? If we trust $\widehat{V}_{\pi,\,2}$ a little more than $\widehat{V}_{\pi,\,1}$, we could consider using

$$\widehat{V}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \left(\widehat{V}_{\pi,\,1}(X_{i1}) + \gamma_{i1}(\pi) \left(\widehat{V}_{\pi,\,2}(X_{i1}, W_{i1}, X_{i2}) - \widehat{V}_{\pi,\,1}(X_{i1}) \right) \right),$$

i.e., on the event where W_{i1} matches π in the first step, we use $\widehat{V}_{\pi,2}$ to debias $\widehat{V}_{\pi,1}$. Here, the γ_{it} are the inverse-propensity weights as in (12.10).

Then next natural question, of course, is why not debias $\widehat{V}_{\pi,2}$ using $\widehat{V}_{\pi,3}$ when W_{i2} also matches π in the second step? And we can do so, and can proceed along until we get to the end of the trajectory, where we interpret $V_{\pi,T+1} = Y$. The expression we get directly by plugging in a doubly robust score to replace $\widehat{V}_{\pi,t}$ is rather unwieldy, but we can rearrange to sum to get

$$\begin{aligned}\widehat{V}_{AIPW}(\pi) = \frac{1}{n}\sum_{i=1}^{n}\Bigg(&\gamma_{iT}Y_i \\ &+ \sum_{t=1}^{T}\left(\gamma_{i(t-1)}(\pi) - \gamma_{it}(\pi)\right)\widehat{V}_{\pi,t}(X_{i1},\,\ldots,\,X_{it})\Bigg),\end{aligned} \tag{12.14}$$

which we recognize as a generalization of the AIPW estimator of Robins, Rotnitzky, and Zhao [1994] for the non-dynamic case.

Bibliographic notes The study of sequential decision rules is a huge topic we've only scratched the surface of. Hernán and Robins [2020] is a good textbook reference. In the statistics literature, a lot of the early results on estimation under sequential ignorability are due to Robins, going back to Robins [1986]. Two helpful references in this line of work are Murphy [2005] and Robins [2004]. The form of the AIPW estimator (12.14) was independently derived by Jiang and Li [2016] and Zhang, Tsiatis, Laber, and Davidian [2013]; see also Thomas and Brunskill [2016].

Lecture 13
Structural Equation Modeling

For most of this class, we've framed causal questions in terms of a potential outcomes model. There is, however, a large alternative tradition of causal analysis based on structural equation modeling. We saw one example of a structural equation model (SEM) in Lecture 9, when introducing instrumental variables methods: We wrote

$$Y_i = \alpha + W_i\tau + \varepsilon_i, \tag{13.1}$$

but did not assume W_i to be uncorrelated with ε_i; we then discussed how an exogenous instrument could be used to identify τ. We called (13.1) a "structural" model because it's not short-hand for an application of least-squares regression; rather, it's a claim that if we had set $W_i = w$, then we would observe an outcome $Y_i(W_i = w) = \alpha + w\tau + \varepsilon_i$.

Our goal today is to survey general results on SEMs: We'll go over how to represent non-parametric SEMs via a directed acyclic graph (DAG), and discuss a general approach to identifying causal effects in such models via the "do calculus." Overall, we'll find that non-parametric SEMs present a powerful and abstract approach to causal inference that sheds new light on familiar identification strategies (e.g., unconfoundedness) and helps unlock some new ones (e.g., the front door criterion). On the other hand, the abstraction of SEMs is less helpful in formalizing some other estimation strategies discussed in class, such as the regression discontinuity design or LATE-focused instrumental variables methods.

Non-parametric SEM Let $(X_1, ..., X_p)$ denote a set of p random variables with joint distribution $\mathbb{P}$, some of which may be observed by the statistician and others not. One can always represent this joint distribution in terms of a

DAG G, meaning that $\mathbb{P}$ factors as

$$\mathbb{P}\left[X_1, ..., X_p\right] = \prod_{j=1}^{p} \mathbb{P}\left[X_j \,\middle|\, pa_j\right], \tag{13.2}$$

where pa_j stands for the parents of X_j in the graph G (i.e., $pa_j = \{X_i : E_{ij} = 1\}$, where E_{ij} denotes the presence of an edge from i to j in G). The decomposition (13.2) becomes a structural model once we further posit the existence of deterministic functions $f_j(\cdot)$, $j = 1, \ldots, p$, such that

$$X_j = f_j\left(pa_j, \varepsilon_j\right), \tag{13.3}$$

where the $\varepsilon_j \sim F_j$ are mutually independent noise terms. The difference between (13.2) and (13.3) is that the former merely characterizes the sampling distribution of X, while the latter lets us also reason about how the value of X_j would change if we were to alter the value of its parents.[1]

Given a SEM (13.3), a causal query involves exogenously setting the values of some nodes of the graph G, and seeing how this affects the distribution of other nodes. Specifically, given two disjoint sets of nodes W, $Y \subset X$, the causal effect of setting W to w on Y is written $\mathbb{P}\left[Y \,\middle|\, do(W = w)\right]$, and corresponds to deleting all equations corresponding to nodes W in (13.3) and plugging in w for W in the rest. In the case where we intervene on a single node X_j, one can check that

$$\mathbb{P}\left[X \,\middle|\, do(X_j = x_j)\right] = \begin{cases} \mathbb{P}\left[X\right] \,/\, \mathbb{P}\left[X_j = x_j \,\middle|\, pa_j\right] & \text{if } X_j = x_j \\ 0 & \text{else.} \end{cases} \tag{13.4}$$

One of the major goals of (non-parametric) structural equation modeling is to provide general methods for answering causal queries in terms of the observed distribution of X using only information provided by the structural model (13.3).[2]

[1] In other words, given (13.2), there always exists a set of functions $f_j(\cdot)$ for which (13.3) when X is drawn according to $\mathbb{P}$. The model (13.3) becomes structural once we assert that the $f_j(\cdot)$ would not change even if we change the sampling distribution of some upstream variables.

[2] Today, we'll never make any functional form assumptions on the model (13.3). For concreteness, you may always assume that X_j is discrete and f_j indexes over distributions for X_j in terms of the values of its parents pa_j. Thus, inference in the linear model (13.1) will not be covered by our discussion today.

The do calculus One nice fact about non-parametric SEM is that there exist powerful abstract tools for reasoning about causal queries. In particular, Pearl [1995] introduced a set of rules, called the do calculus, which lets us verify whether causal queries are answerable in terms of the graph G underlying (13.3).

To understand do calculus, it is helpful to first recall how graphs encode conditional independence statements in terms of d-separation, defined as follows. Let X, Y and Z denote disjoint sets of nodes, and let p be any (undirected) path from a node in X to a node in Y. We say that Z blocks p if there is a node W on p such that either (i) W is a collider on p (i.e., W has two incoming edges along p) and neither W nor any of its descendants are in Z, or (ii) W is not a collider and W is in Z. We say that Z d-separates X and Y if it blocks every path between X and Y.

The motivation behind this definition is that, as shown by Geiger, Verma, and Pearl [1990], d-separation encodes every conditional independence statement implied by the graph factorization (13.2), i.e., we can deduce $X \perp\!\!\!\perp Y \,\big|\, Z$ from (13.2) if and only if Z d-separates X and Y in the graph G. Motivated by this fact, we write d-separation as $(X \perp\!\!\!\perp Y \,\big|\, Z)_G$.

Do calculus provides a way to simplify causal queries by referring to d-separation on various sub-graphs of G. To this end define $G_{\overline{X}}$ the subgraph of G with all edges incoming to X deleted, $G_{\underline{X}}$ the subgraph of G with all outgoing edges from X deleted, $G_{\underline{X}\overline{Z}}$ the subgraph of G with all outgoing edges from X and incoming edges to Z deleted, etc. Then, for any disjoint sets of edges X, Y, Z, W the following equivalence statements hold.

1. Insertion/deletion of observations: If $\left(Y \perp\!\!\!\perp Z \,\big|\, X,\, W\right)_{G_{\overline{X}}}$ then

$$\begin{aligned} &\mathbb{P}\left[Y \,\big|\, do(X = x),\, Z = z,\, W = w\right] \\ &\qquad = \mathbb{P}\left[Y \,\big|\, do(X = x),\, W = w\right]. \end{aligned} \tag{13.5}$$

2. Action/observation exchange: If $\left(Y \perp\!\!\!\perp Z \,\big|\, X,\, W\right)_{G_{\overline{X}\underline{Z}}}$ then

$$\begin{aligned} &\mathbb{P}\left[Y \,\big|\, do(X = x),\, do(Z = z),\, W = w\right] \\ &\qquad = \mathbb{P}\left[Y \,\big|\, do(X = x),\, Z = z,\, W = w\right]. \end{aligned} \tag{13.6}$$

3. Insertion/deletion of actions: If $\left(Y \perp\!\!\!\perp Z \,\big|\, X,\, W\right)_{G_{\overline{X}\overline{Z(W)}}}$ where $Z(W)$ is the set of Z nodes that are not ancestors of any W node in $G_{\overline{X}}$, then

$$\begin{aligned} &\mathbb{P}\left[Y \,\big|\, do(X = x),\, do(Z = z),\, W = w\right] \\ &\qquad = \mathbb{P}\left[Y \,\big|\, do(X = x),\, W = w\right]. \end{aligned} \tag{13.7}$$

When applying the do calculus, our goal is to apply these 3 rules of inference until we've reduced a causal query to a query about observable moments of $\mathbb{P}$, i.e., conditional expectations that do not involve the do-operator and that only depend on observed random variables. As shown in subsequent work, the do calculus is complete, i.e., if we cannot use the do calculus to simply a causal query then it is not non-parametrically identified in terms of the structural equation model; see Pearl [2009] for a discussion and references.

Example 1: Back-door criterion Suppose have disjoint sets of nodes X, Y, W, and want to query $\mathbb{P}\left[Y \,\middle|\, do(W = w)\right]$. Suppose moreover that X contains no nodes that are downstream for W, and that X d-separates W and Y once we block all downstream edges from W, i.e., that

$$\left(Y \perp\!\!\!\perp W \,\middle|\, X\right)_{G_{\underline{W}}}. \tag{13.8}$$

Then, we can identify the effect of W on Y via

$$\mathbb{P}\left[Y \,\middle|\, do(W = w)\right] = \sum_x \mathbb{P}\left[X = x\right] \mathbb{P}\left[Y \,\middle|\, X = x,\, W = w\right]. \tag{13.9}$$

To verify (13.9), we can use the rules of do calculus as follows:

$$\begin{aligned}
\mathbb{P}\left[Y \,\middle|\, do(W = w)\right] &= \sum_x \mathbb{P}\left[X = x \,\middle|\, do(W = w)\right] \mathbb{P}\left[Y \,\middle|\, X = x,\, do(W = w)\right] \\
&= \sum_x \mathbb{P}\left[X = x\right] \mathbb{P}\left[Y \,\middle|\, X = x,\, do(W = w)\right] \\
&= \sum_x \mathbb{P}\left[X = x\right] \mathbb{P}\left[Y \,\middle|\, X = x,\, W = w\right],
\end{aligned}$$

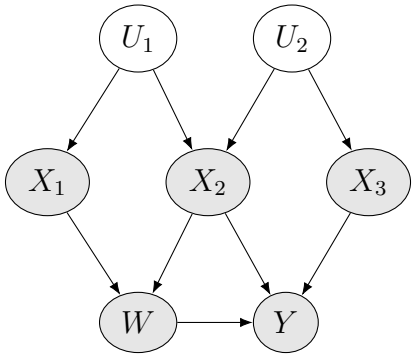
where the first equality is just the chain rule, the second equality follows from rule #3 because X is upstream from W and so $(X \perp\!\!\!\perp W)_{G_{\overline{W}}}$, and the third equality follows from rule #2 by (13.8).

The back-door criterion is of course closely related to unconfoundedness, and the identification strategy (13.9) exactly matches the standard regression adjustment under unconfoundedness. To understand the connection between (13.8) and unconfoundedness, consider the case where Y and W are both singletons and W has no other downstream variables in G other than Y. Then, blocking downstream arrows from W can be interpreted as leaving the effect of W on Y unspecified, and (13.8) becomes

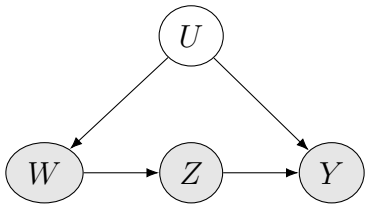
$$F_Y(w) \perp\!\!\!\perp W \,\middle|\, X, \tag{13.10}$$

where $F_Y(w) = f_Y(w, X, \varepsilon_Y)$ leaves all but the contribution of w unspecified in (13.3). The condition is clearly analogous to unconfoundedness (although the fundamental causal model is different).

One useful consequence of this back-door criterion result is that we can now reason about the main conditional independence condition (13.8) via the graphical d-separation rule. Consider, for example, the graph below. By applying d-separation above, one immediately sees that (13.8) holds if we condition on $\{X_1, X_2\}$ or $\{X_2, X_3\}$, but not if we only condition on X_2. In contrast, the classical presentation based on unconfoundedness asks the scientist to simply assert a conditional independence statement of the type (13.10), and does not provide tools like d-separation that could be used to reason about when such a condition might hold in the context of slightly more complicated stochastic models.



Example 2: Front-door criterion Another application of do calculus that results in something much less familiar arises in the following graph. We still want to compute $\mathbb{P}\left[Y \,\middle|\, do(W = w)\right]$, but now do not observe U and so cannot apply the backdoor criterion. However, if there exists a variable Z which, like in the graph below, fully mediates the effect of W on Y without being affected by U, we can use it for identification.



We proceed as follows. First, following the same line of argumentation as

before, we see that

$$\begin{aligned}\mathbb{P}\left[Y \,\middle|\, do(W=w)\right] &= \sum_z \mathbb{P}\left[Z=z \,\middle|\, do(W=w)\right] \mathbb{P}\left[Y \,\middle|\, Z=z,\, do(W=w)\right] \\ &= \sum_z \mathbb{P}\left[Z=z \,\middle|\, W=w\right] \mathbb{P}\left[Y \,\middle|\, Z=z,\, do(W=w)\right],\end{aligned}$$

where the first equality is the chain rule and the second equality is from the back-door. We have to work a little harder to resolve the second term, however. Here, the main idea is to start by taking one step backwards before proceeding further:

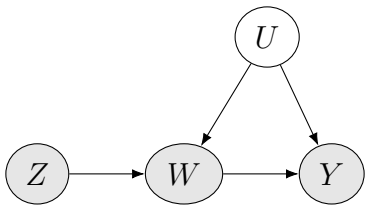
$$\begin{aligned}\mathbb{P}\left[Y \,\middle|\, Z=z,\, do(W=w)\right] &= \mathbb{P}\left[Y \,\middle|\, do(Z=z),\, do(W=w)\right] \\ &= \mathbb{P}\left[Y \,\middle|\, do(Z=z)\right] \\ &= \sum_{w'} \mathbb{P}\left[W=w'\right] \mathbb{P}\left[Y \,\middle|\, Z=z,\, W=w'\right],\end{aligned}$$

where the first equality follows from rule #2, the second equality follows from rule #3, and the last is just the backdoor adjustment again. Plugging this in, we find that

$$\begin{aligned}&\mathbb{P}\left[Y \,\middle|\, do(W=w)\right] \\ &\quad = \sum_z \mathbb{P}\left[Z=z \,\middle|\, W=w\right] \sum_{w'} \mathbb{P}\left[W=w'\right] \mathbb{P}\left[Y \,\middle|\, Z=z,\, W=w'\right]. \end{aligned} \tag{13.11}$$

This identification formula is often called the front-door criterion. Interestingly, even though it queries about a $do(W=w)$ intervention, it still integrates over the observed distribution of $\mathbb{P}\left[W=w'\right]$.

Example 3: Instrumental variables A last setting of interest, pictured below, represents the instrumental variables setting. We want to estimate $\mathbb{P}\left[Y \,\middle|\, do(W=w)\right]$, and there's an unobserved confounder U that prevents us from applying the back-door criterion:

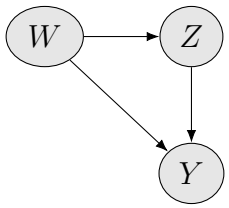


We'd want to use an instrument Z that exogenously nudges W for identification. Here, however, do calculus does not help us: There's no way to identify $\mathbb{P}\left[Y \,\middle|\, do(W = w)\right]$ in the graph below. And, in fact, there's no way something like an instrument could ever help us: Adding more nodes to a graph just makes it strictly harder to satisfy the d-separation conditions used for do calculus.

What went wrong here? Note that, when discussing IV methods in previous lectures, we had to be very careful with our assumptions. We typically needed to assume something more than just the SEM above (e.g., monotonicity of response to the instrument) and, even so, we could usually only identify non-standard causal quantities like the local average treatment effect—and we used carefully crafted relationships between potential outcomes to express these facts. In contrast, embedding further assumptions like monotonicity into SEM appears challenging.

SEMs and potential outcomes Structural equation modeling and potential outcome modeling are of course closely related, and share the same overarching goal. They both let us reason about how exogenously changing one variable could affect others. And, in a simple two-node graph $W \to Y$, they are in fact the same: One can create a one-to-one mapping between potential outcomes $Y(w)$ and the SEM representation $f_Y(w, \varepsilon_Y)$.

More generally, however, the SEM and potential outcomes formalisms do not match. When working with potential outcomes, all causal effects correspond to specific manipulations, i.e., we cannot even ask causal questions that aren't of the form "what would Y be if I experimentally set these other variables to specific values" [Holland, 1986]. In contrast, SEM lets us ask causal questions that do not reduce to manipulations. Consider the following simple DAG:



Furthermore, write $z(w) := f_Z(w, \varepsilon_Z)$ for the value Z would have taken had we set W to w. With SEMs, nothing stop us from querying about objects like

$$\begin{aligned}\mathbb{P}\left[Y \,\middle|\, do(W = 0),\, do(Z = z(w = 1))\right] \\ = f_Y\left(w = 0,\, z = f_Z(w = 1,\, \varepsilon_Z),\, \varepsilon_Y\right).\end{aligned} \tag{13.12}$$

On the other hand, in the potential outcomes setting, queries of this type don't really make sense (did we set W to 0 or 1?), unless we augment the graph with more nodes so that the specified query corresponds to a well defined manipulation in the augmented graph.

Questions of this type have led to considerable debate in the causal inference community. Some proponents of potential outcomes modeling argue that the ability of SEM to identify causal effects that do not correspond to manipulations is a flaw of the SEM framework. On the other hand, Pearl argues that the inability of the potential outcomes framework to express non-manipulable causal effects is a limitation of that framework.

Bibliographic notes The do calculus was proposed by Pearl [1995], and today's notes in large part follow the exposition of that paper, including the examples of the front- and back-door criteria. A recent overview of the corresponding literature is given in Pearl [2009]. One should note that structural equation models are not the only way of representing causal effects in complex sampling designs using DAGs; other approaches have also been developed by Robins [1986] and Spirtes, Glymour, and Scheines [1993]. In particular, the approach of Robins [1986] builds on the potential outcomes framework, and thus does not allow us to reason able non-manipulable causes; see Robins and Richardson [2010] for further discussion. For a broader discussion of the role of SEMs in empirical work, see Imbens [2019], Pearl and Mackenzie [2018], and references therein.

Lecture 14
Adaptive Experiments

So far, we've mostly focused on data collected in an IID setting. In the context of a randomized trial, the IID setting involves pre-committing to experiment on a certain number of study participants, and assigning each of them to treatment with a certain pre-specified probability. In many settings, however, the structure of an IID randomized trial may be too rigid. For example, in a vaccine trial, if we notice that some candidate vaccines are failing to produce antibodies, then we may want to eliminate them from the study (for both cost and ethical reasons).

Today, we'll survey some results on the design of adaptive experiments, which enable the analyst to shift their data collection scheme in response to preliminary findings. There is a wide variety of methods (often called bandit algorithms), that can be used to accomplish this task. Our goal is to review how such adaptive experimentation schemes can be used to mitigate the cost of having bad intervention arms in our experiment, and also discuss some challenges in doing inference with adaptively collected data.

Setting and notation We are interested in studying the relative value of $k = 1, \ldots, K$ candidate actions, and to do so have access to a stream of $t = 1, \ldots, T$ experimental subjects. Each subject has IID potential outcomes, and we observe the potential outcome corresponding to our action,

$$Y_t(k) \sim F_k, \quad Y_t = Y_t(W_t), \tag{14.1}$$

where W_t is the action taken at time t and F_k is the potential outcome distribution for the k-th arm. We write $\mu_k = \mathbb{E}\left[Y_t(k)\right]$ for the mean of F_k, and define regret as

$$R_T = \sum_{t=1}^{T} \left(\mu^* - \mu_{W_t}\right), \quad \mu^* = \sup\left\{\mu_k : 1 \leq k \leq K\right\} \tag{14.2}$$

as the (expected) shortfall in rewards given our sequence of actions. Throughout, we write

$$n_{k,t} = \sum_{j=1}^{t} 1\left(\{W_j = k\}\right), \quad \hat{\mu}_{k,t} = \frac{1}{n_{k,t}} \sum_{j=1}^{t} 1\left(\{W_j = k\}\right) Y_j \tag{14.3}$$

for the cumulative number of times the k-th arm has been drawn and the current running average of rewards from it. Clearly, in a randomized trial where W_t is uniformly (and non-adaptively) distributed on $\{1, \ldots, K\}$, regret scales linearly in T, i.e., $R_T \sim T \sum_{k=1}^{K} (\mu^* - \mu_k) / K$.

A first goal of adaptive experimentation schemes is to do better, and achieve sub-linear regret. In order to do so, any algorithm will first need to explore the sampling distribution to figure out which arms $k = 1, \ldots, K$ are the most promising, and then exploit this knowledge to attain low regret.

Optimism in the face of uncertainty One notable early solution to the explore-exploit trade-off problem in adaptive experiments in the upper confidence band (UCB) algorithm of Lai and Robbins [1985]. The algorithm proceeds as follows. First, initialize each arm using t_0 draws and then,

- At each time $t = Kt_0 + 1$, $Kt_0 + 2$, ..., construct a confidence interval $\widehat{U}_{k,t}$ for μ_k based on data collected up to time $t - 1$, and
- Pick action W_t corresponding to the confidence interval $\widehat{U}_{k,t}$ with the largest upper endpoint, and observe $Y_t = Y_t(W_t)$.

At a high level, the motivation behind UCB is that we always want to explore the arm with the most upside. At the beginning of time we have a lot of uncertainty about each arm, and so we optimistically sample all of them. Over time, however, we'll collect enough data from the bad arms to be fairly sure they're suboptimal, and at that point UCB will start sampling them less. There are many different variants of UCB considered in practice that arise from different constructions for the confidence interval $\widehat{U}_{k,t}$ used for arm selection.

To get an understanding of why UCB controls regret, consider a simplification of the sampling model (14.1) with Gaussian F_k, i.e.,

$$Y_t(k) \sim \mathcal{N}\left(\mu_k, \sigma^2\right), \tag{14.4}$$

where σ^2 is known. We run UCB with confidence intervals[1]

$$\widehat{U}_{k,t} = \hat{\mu}_{k,t-1} \pm \sigma \sqrt{4 \log(T) \,/\, n_{k,t-1}}. \tag{14.5}$$

[1]Here, the Gaussianity and known σ and T assumptions help simplify the proof; one can get rid of them at the expense of a slightly more delicate algorithm and argument.

One can then verify the following. Under our sampling assumptions, UCB with intervals (14.5) and $t_0 = 1$ initial draws has regret bounded as

$$R_T = \sum_{k \neq k^*} \frac{16\sigma^2 \log(T)}{\mu_{k^*} - \mu_k} + (\mu_{k^*} - \mu_k) \quad \text{with prob. at least} \quad 1 - K/T, \tag{14.6}$$

where k^* denotes the optimal arm. This result immediately implies that UCB in fact succeeds in finding and effectively retiring sub-optimal arms reasonably fast, thus resulting in regret that only scales logarithmically in the regret. Interestingly, the dominant term in (14.6) is due to "good" arms for which $\mu^* - \mu_k$ is small; intuitively, the reason these arms are difficult to work with is that it takes longer to be sure that they're sub-optimal. This implies that the cost of including some really bad arms in an adaptive experiment may be limited, since an algorithm like UCB will be able to discard them quickly.

Finally, one should note that the upper bound (14.6) appears to allow for unbounded regret due to quasi-optimal arms for which $\mu_{k^*} - \mu_k$ is very small. This is simply an artifact of the proof strategy, that focused on the case where effects are strong. When effects may be weak, one can simply note that the worst-case regret due to any given arm k is upper bounded by $T(\mu_{k^*} - \mu_k)$; and, combining this bound with the bound implied by (14.6), we find that the worst-case regret for any combination of arms μ_k is bounded on the order of $K\sqrt{T \log(T)}$.

A regret bound for UCB In order to prove (14.6), we first note that regret R_T can equivalently be expressed as

$$R_T = \sum_{k \neq k^*} n_{k,T} (\mu_{k^*} - \mu_k). \tag{14.7}$$

Our main task is thus to bound $n_{k,T}$, i.e., the number of times UCB may pull any sub-optimal arm; and it turns out that UCB is essentially an algorithm reverse-engineered to make such an argument go through. To this end, the first thing to check is that, with probability $1 - 1/T$ and every arm $k = 1, ..., K$, we have

$$\mu_k \leq \hat{\mu}_{k,t-1} + \sigma \sqrt{4 \log(T) \,/\, n_{k,t-1}} \tag{14.8}$$

for all $t = K+1, \ldots, T$. This is true because, writing $\zeta_{k,j}$ for the j-th time arm k was pulled, we have

$$\begin{aligned}
&\mathbb{P}\left[\sup_{K<t\leq T}\left\{\mu_k - \hat{\mu}_{k,t-1} - \sigma\sqrt{4\log(T)\,/\,n_{k,t-1}} \geq 0\right\}\right] \\
&\quad\leq \mathbb{P}\left[\sup_{1\leq j\leq n_{k,T}}\left\{\mu_k - \hat{\mu}_{k,\zeta_{k,j}} - \sigma\sqrt{4\log(T)\,/\,j} \geq 0\right\}\right] \\
&\quad= \mathbb{P}\left[\sup_{1\leq j\leq n_{k,T}}\left\{\mu_k - \frac{1}{j}\sum_{l=1}^{j} Y_l'(0) - \sigma\sqrt{4\log(T)\,/\,j} \geq 0\right\}\right] \\
&\quad\leq \mathbb{P}\left[\sup_{1\leq j\leq T}\left\{\mu_k - \frac{1}{j}\sum_{l=1}^{j} Y_l'(0) - \sigma\sqrt{4\log(T)\,/\,j} \geq 0\right\}\right] \\
&\quad\leq 1/T,
\end{aligned}$$

where the equality follows by stationarity of the data-generating process (here, $Y_l'(k)$ are independent draws from $\mathcal{N}(\mu_k, \sigma^2)$), and the last line is an application of Hoeffding's inequality together with a union bound. By another union bound, we see that (14.8) holds for all bounds with probability at least $1-K/T$.

Then, on the event where (14.8) holds for all arms, we see that we can only pull arm k under the following (necessary but not sufficient) conditions, where k^* denotes the optimal arm:

$$\begin{aligned}
W_t = k &\implies \hat{\mu}_{k,t-1} + \sigma\sqrt{4\log(T)\,/\,n_{k,t-1}} \geq \hat{\mu}_{k^*,t-1} + \sigma\sqrt{4\log(T)\,/\,n_{k^*,t-1}} \\
&\implies \hat{\mu}_{k,t-1} + \sigma\sqrt{4\log(T)\,/\,n_{k,t-1}} \geq \mu_{k^*} \\
&\implies \mu_k + 2\sigma\sqrt{4\log(T)\,/\,n_{k,t-1}} \geq \mu_{k^*} \\
&\implies n_{k,t-1} \leq 16\sigma^2\log(T)/\left(\mu_{k^*} - \mu_k\right)^2.
\end{aligned}$$

Thus, when (14.8) holds for all arms, pulling the k-th arm simply becomes impossible once $n_{k,t-1}$ passes a certain cutoff. Plugging this bound on n_k, T into the regret expression

$$R_T = \sum_{k\neq k^*}\left(\mu_{k^*} - \mu_k\right) n_{k,T}, \tag{14.9}$$

we obtain (14.6).

Adaptive randomization schemes UCB is a simple approach to adaptive experimentation with strong bounds on excess regret from sampling suboptimal arms. However, from a practical point of view, it has some fairly

serious limitations. First, and most importantly, UCB cannot be understood as an adaptive randomized experiment: The choice of action W_t is a deterministic function of past observations. This makes it difficult to interface between UCB and methods designed for the IID setting that explicitly rely on randomization. Second, one might be qualitatively concerned that the form of the UCB algorithm is too closely linked to the proof technique, and this may make it difficult to generalize UCB to more complicated sampling designs[2]

One popular alternative to UCB that helps address these issues is Thompson sampling [Thompson, 1933]. Thompson sampling is a heuristic algorithm based on Bayesian updating. To start, we pick a prior $\Pi_{k,0}$ on the potential outcome distribution F_k in (14.1). Then, for each time $t = 1, \ldots, T$, we

- Compute probabilities $e_{k,t-1}$ that each arm k is the best arm, i.e.,

$$e_{k,t-1} = \mathbb{P}_{\Pi_{\cdot,t-1}}\left[\mu_k = \mu_*\right], \tag{14.10}$$

- Randomly choose an action $W_t \sim \text{Multinomial}(e_{\cdot,t-1})$, and
- Observe $Y_t = Y_t(W_t)$ and update the posterior $\Pi_{\cdot,t}$.

Although Thompson sampling looks superficially very different from UCB, it ends up having a very similar statistical behavior to it. Just like UCB, Thompson sampling regularly explores every arm until it becomes effectively sure that the arm is not good (i.e., the posterior probability of the arm being best drops below $1/T$); and intuition from, say, the Bernstein–von Mises theorem suggests that this should happen with roughly the same amount of information as when the upper confidence band of an arm falls below the whole confidence interval of some better arm.

Methodologically, meanwhile, Thompson sampling presents a rather desirable alternative to UCB. The actions taken during Thompson sampling are randomized (with adaptive randomization probabilities that depend on past data), thus opening the door to a tighter connection with the causal inference literature. And the main tuning parameter in Thompson sampling, namely the set of priors $\Pi_{\cdot,0}$, is often easier to reason about in practice than the choice of confidence band construction for UCB.

[2]Even in the slightly more general case where σ_k^2 is unknown and may vary across arms, one needs to adapt the form of the UCB confidence intervals (14.5) so that they allow for a different proof that builds on different concentration inequalities, and there are several choices for how to do so.

Inference in adaptive experiments Both UCB an Thompson sampling provide powerful approaches to adaptive data collection that don't occur much regret even when some very bad arms may be initially under consideration. However, once we've collected this data, we will often want to analyze it and, e.g., provide confidence statements for the underlying problem parameters. Such analysis, however, is considerably more difficult than in the IID setting. For example, in the case of estimating μ_k, two natural estimators that immediately come to mind include the sample mean

$$\hat{\mu}_k^{AVG} = \hat{\mu}_{k,T} = \frac{1}{n_{k,T}^{-1}} \sum_{j=1}^{t} 1\left(\{W_j = k\}\right) Y_j \tag{14.11}$$

and, in the case of Thompson sampling, the inverse-propensity weighted estimator

$$\hat{\mu}_k^{IPW} = \frac{1}{T} \sum_{t=1}^{T} \frac{1\left(\{W_t = k\}\right) Y_t}{e_{t,k}}. \tag{14.12}$$

However, due to the adaptive data-collection scheme, neither of these estimators has an asymptotically normal limiting distribution, thus hindering their use for making confidence intervals.

Perhaps surprisingly, however, it's possible to design adaptively weighted estimates of μ_k that do admit a Gaussian pivot. One example of such weighting scheme is

$$\hat{\mu}_k^{AW} = \sum_{t=1}^{T} \frac{1\left(\{W_t = k\}\right) Y_t}{\sqrt{e_{t,k}}} \Big/ \sum_{t=1}^{T} \frac{1\left(\{W_t = k\}\right)}{\sqrt{e_{t,k}}}, \tag{14.13}$$

which, under reasonable regularity conditions, satisfies

$$\begin{aligned} &\widehat{V}_k^{-1/2}\left(\hat{\mu}_k^{AW} - \mu_k\right) \Rightarrow \mathcal{N}\left(0,\, 1\right), \\ &\widehat{V}_k = \sum_{t=1}^{T} \left(\frac{1\left(\{W_t = k\}\right)\left(Y_t - \hat{\mu}_k^{AW}\right)}{\sqrt{e_{t,k}}} \right)^2 \Big/ \left(\sum_{t=1}^{T} \frac{1\left(\{W_t = k\}\right)}{\sqrt{e_{t,k}}} \right)^2. \end{aligned} \tag{14.14}$$

The reason these weights help restore a CLT is that they are "variance stabilizing", meaning that the variance of the resulting estimator is predictable in the sense required by relevant central limit theorems. To verify (14.13), we note that

$$\hat{\mu}_k^{AW} - \mu_k = \sum_{t=1}^{T} \frac{1\left(\{W_t = k\}\right)\left(Y_t - \mu_k\right)}{\sqrt{e_{t,k}}} \Big/ \sum_{t=1}^{T} \frac{1\left(\{W_t = k\}\right)}{\sqrt{e_{t,k}}}, \tag{14.15}$$

and start by focusing on the numerator of the above expression. Let

$$M_t = \sum_{j=1}^{t} \frac{1\left(\{W_j = k\}\right)\left(Y_j - \mu_k\right)}{\sqrt{e_{j,k}}} \tag{14.16}$$

be its partial sum. Because W_t is randomly chosen given information up to time t, we see that W_t is independent of $Y_t(k)$ conditionally on $M_{1:(t-1)}$, and thus M_t is a martingale:

$$\mathbb{E}\left[M_t \,\middle|\, M_{1:(t-1)}\right] = M_{t-1}. \tag{14.17}$$

Furthermore, thanks to our weighting scheme, we can check that

$$\mathrm{Var}\left[M_t \,\middle|\, M_{1:(t-1)}\right] = \sigma_k^2, \qquad \sigma_k^2 = \mathrm{Var}\left[Y_t(k)\right]. \tag{14.18}$$

Given these two facts, one can use a martingale central limit theorem as given in, e.g., Helland [1982], that provided the $e_{t,k}$ do not decay too fast,[3]

$$M_T \,/\, \sqrt{T\sigma_k^2} \Rightarrow \mathcal{N}\left(0,\, 1\right). \tag{14.19}$$

The central limit theorem (14.14) follows by noting that

$$\sum_{t=1}^{T} \left(\frac{1\left(\{W_t = k\}\right)\left(Y_t - \hat{\mu}_k^{AW}\right)}{\sqrt{e_{t,k}}} \right)^2 \Big/ \left(T\sigma_k^2\right) \to_p 1 \tag{14.20}$$

by martingale concentration (again provided the propensities don't decay too fast), and that the denominators of $\hat{\mu}_k^{AW}$ and $\widehat{V}_k^{1/2}$ cancel out in (14.14). The key step in this proof that would not have held for alternative estimators (such as the unweighted sample mean) is (14.18).

Bibliographic notes This line of work on bandit algorithms builds on early results from Lai and Robbins [1985] on the UCB algorithm. Lai and Robbins [1985] showed that a variant of UCB achieves regret scaling of the form (14.6), and that this behavior is asymptotically optimal. A more recent analysis of UCB without parametric assumptions on the reward distribution F_k is given in Auer, Cesa-Bianchi, and Fischer [2002], while Agrawal and Goyal [2017] provide analogous bounds for Thompson sampling. Thanks to its Bayesian

[3]One tension here is that, in general, adaptively weighted CLTs require the sampling probabilities $e_{t,k}$ to decay slower that $1/t$, which rules out sampling schemes that get the optimal $\log(T)$ regret with strong signals.

specification, Thompson sampling can be generalize to a wide variety of adaptive learning problems; see Russo, Van Roy, Kazerouni, Osband, and Wen [2018] for a recent survey.

The line of work on inference with adaptively collected data via variance-stabilizing weighting is pursued by Luedtke and van der Laan [2016] and Hadad, Hirshberg, Zhan, Wager, and Athey [2019]. One should note that this is not the only possible approach to inference in adaptive experiments. In particular, a classical alternative to inference in this setting starts from confidence-bands based on the law of the iterated logarithm and its generalizations that hold simultaneously for every value of t; see Robbins [1970] for a landmark survey and Howard, Ramdas, McAuliffe, and Sekhon [2018] for recent advances.

Finally, all approaches to adaptive experimentation discussed today are essentially heuristic algorithms that can be shown to have good asymptotic behavior (i.e., neither UCB nor Thompson sampling can be derived directly from an optimality principle). In the Bayesian case (i.e., where we have an actual subjective prior for F_k rather than just a convenience prior as used by Thompson sampling to power an algorithm with frequentist guarantees), it is possible to solve for the optimal regret-minimizing experimental design via dynamic programming [Gittins, 1979].

Bibliography

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

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

Alberto Abadie and Guido W Imbens. Matching on the estimated propensity score. *Econometrica*, 84(2):781–807, 2016.

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of californias tobacco control program. *Journal of the American Statistical Association*, 105(490):493–505, 2010.

Shipra Agrawal and Navin Goyal. Near-optimal regret bounds for Thompson sampling. *Journal of the ACM*, 64(5):1–24, 2017.

Luigi Ambrosio and Gianni Dal Maso. A general chain rule for distributional derivatives. *Proceedings of the American Mathematical Society*, 108(3):691–702, 1990.

Takeshi Amemiya. The nonlinear two-stage least-squares estimator. *Journal of Econometrics*, 2(2):105–110, 1974.

Joshua D Angrist, Guido W Imbens, and Donald B Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996.

Joshua D Angrist, Kathryn Graddy, and Guido W Imbens. The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish. *The Review of Economic Studies*, 67 (3):499–527, 2000.

Manuel Arellano. *Panel Data Econometrics*. Oxford university press, 2003.

Dmitry Arkhangelsky and Guido W Imbens. Double-robust identification for causal panel data models. *arXiv preprint arXiv:1909.09412*, 2019.

Dmitry Arkhangelsky, Susan Athey, David A Hirshberg, Guido W Imbens, and Stefan Wager. Synthetic difference in differences. *arXiv preprint arXiv:1812.09970*, 2018.

Timothy B Armstrong and Michal Kolesár. Optimal inference in a class of regression models. *Econometrica*, 86(2):655–683, 2018.

Timothy B Armstrong and Michal Kolesár. Simple and honest confidence intervals in nonparametric regression. *Quantitative Economics*, 11(1):1–39, 2020.

Susan Athey and Guido W Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27): 7353–7360, 2016.

Susan Athey and Stefan Wager. Efficient policy learning. *arXiv preprint arXiv:1702.02896*, 2017.

Susan Athey and Stefan Wager. Estimating treatment effects with causal forests: An application. *Observational Studies*, 5:36–51, 2019.

Susan Athey, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi. Matrix completion methods for causal panel data models. *arXiv preprint arXiv:1710.10251*, 2017.

Susan Athey, Guido W Imbens, and Stefan Wager. Approximate residual balancing: Debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):597–623, 2018.

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002.

Jushan Bai. Panel data models with interactive fixed effects. *Econometrica*, 77(4):1229–1279, 2009.

Alexandre Belloni, Daniel Chen, Victor Chernozhukov, and Christian Hansen. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429, 2012.

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650, 2014.

Marianne Bertrand, Esther Duflo, and Sendhil Mullainathan. How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics*, 119(1):249–275, 2004.

Dimitris Bertsimas and Nathan Kallus. From predictive to prescriptive analytics. *Management Science*, 66(3):1025–1044, 2020.

Peter J Bickel, Chris AJ Klaassen, Ya'acov Ritov, and Jon A Wellner. *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins University Press Baltimore, 1993.

Adam Bloniarz, Hanzhong Liu, Cun-Hui Zhang, Jasjeet S Sekhon, and Bin Yu. Lasso adjustments of treatment effect estimates in randomized experiments. *Proceedings of the National Academy of Sciences*, 113(27):7383–7390, 2016.

Stéphane Bonhomme and Elena Manresa. Grouped patterns of heterogeneity in panel data. *Econometrica*, 83(3):1147–1184, 2015.

John Bound, David A Jaeger, and Regina M Baker. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association*, 90(430):443–450, 1995.

Andreas Buja, Lawrence Brown, Richard Berk, Edward George, Emil Pitkin, Mikhail Traskin, Kai Zhang, and Linda Zhao. Models as approximations I: Consequences illustrated with linear regression. *Statistical Science*, 34(4): 523–544, 2019.

Sebastian Calonico, Matias D Cattaneo, and Rocio Titiunik. Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, 82(6):2295–2326, 2014.

Sebastian Calonico, Matias D Cattaneo, and Max H Farrell. On the effect of bias estimation on coverage accuracy in nonparametric inference. *Journal of the American Statistical Association*, 113(522):767–779, 2018.

Sebastian Calonico, Matias D Cattaneo, Max H Farrell, and Rocio Titiunik. Regression discontinuity designs using covariates. *Review of Economics and Statistics*, 101(3):442–451, 2019.

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.

Gary Chamberlain. Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, 34(3):305–334, 1987.

Ming-Yen Cheng, Jianqing Fan, and James S Marron. On automatic boundary corrections. *The Annals of Statistics*, 25(4):1691–1708, 1997.

Victor Chernozhukov, Juan Carlos Escanciano, Hidehiko Ichimura, Whitney K Newey, and James M Robins. Locally robust semiparametric estimation. *arXiv preprint arXiv:1608.00033*, 2016.

Victor Chernozhukov, Mert Demirer, Esther Duflo, and Iván Fernández-Val. Generic machine learning inference on heterogenous treatment effects in randomized experiments. *arXiv preprint arXiv:1712.04802*, 2017.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):1–68, 2018a.

Victor Chernozhukov, Whitney Newey, and James Robins. Double/de-biased machine learning using regularized Riesz representers. *arXiv preprint arXiv:1802.08667*, 2018b.

Richard K Crump, V Joseph Hotz, Guido W Imbens, and Oscar A Mitnik. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199, 2009.

Clément de Chaisemartin and Xavier D'Haultfoeuille. Two-way fixed effects estimators with heterogeneous treatment effects. *arXiv preprint arXiv:1803.08807*, 2018.

Alexis Diamond and Jasjeet S Sekhon. Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics*, 95(3):932–945, 2013.

Peng Ding, Avi Feller, and Luke Miratrix. Decomposing treatment effect variation. *Journal of the American Statistical Association*, 114(525):304–317, 2019.

David L Donoho. Statistical estimation and optimal recovery. *The Annals of Statistics*, 22(1):238–270, 1994.

Dean Eckles, Nikolaos Ignatiadis, Stefan Wager, and Han Wu. Noise-induced randomization in regression discontinuity designs. *arXiv preprint arXiv:2004.09458*, 2020.

Bradley Efron. *The Jackknife, the Bootstrap, and other Resampling Plans*. Siam, 1982.

Max H Farrell. Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, 189(1):1–23, 2015.

Constantine E Frangakis and Donald B Rubin. Principal stratification in causal inference. *Biometrics*, 58(1):21–29, 2002.

Dan Geiger, Thomas Verma, and Judea Pearl. Identifying independence in bayesian networks. *Networks*, 20(5):507–534, 1990.

John C Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(2):148–164, 1979.

Bryan S Graham, Cristine Campos de Xavier Pinto, and Daniel Egel. Inverse probability tilting for moment condition models with missing data. *The Review of Economic Studies*, 79(3):1053–1079, 2012.

Trygve Haavelmo. The statistical implications of a system of simultaneous equations. *Econometrica*, 11(1):1–12, 1943.

Vitor Hadad, David A Hirshberg, Ruohan Zhan, Stefan Wager, and Susan Athey. Confidence intervals for policy evaluation in adaptive experiments. *arXiv preprint arXiv:1911.02768*, 2019.

Jinyong Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66(2):315–331, 1998.

Jinyong Hahn, Petra Todd, and Wilbert van der Klaauw. Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, 69(1):201–209, 2001.

P Richard Hahn, Jared S Murray, and Carlos M Carvalho. Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. *Bayesian Analysis*, 2020.

Jens Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20(1):25–46, 2012.

Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC press, 2015.

James J Heckman. Sample selection bias as a specification error. *Econometrica*, 47(1):153–161, 1979.

James J Heckman and Edward J Vytlacil. Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proceedings of the National Academy of Sciences*, 96(8):4730–4734, 1999.

James J Heckman and Edward J Vytlacil. Structural equations, treatment effects, and econometric policy evaluation. *Econometrica*, 73(3):669–738, 2005.

Inge S Helland. Central limit theorems for martingales with discrete or continuous time. *Scandinavian Journal of Statistics*, pages 79–94, 1982.

Miguel A Hernán and James M Robins. *Causal Inference: What If*. Chapman & Hall/CRC, Boca Raton, 2020.

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.

David A Hirshberg and Stefan Wager. Augmented minimax linear estimation. *arXiv preprint arXiv:1712.00038*, 2017.

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

Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Uniform, nonparametric, non-asymptotic confidence sequences. *arXiv preprint arXiv:1810.08240*, 2018.

Kosuke Imai and Marc Ratkovic. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1): 443–470, 2013.

Kosuke Imai and Marc Ratkovic. Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1): 243–263, 2014.

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. Instrumental variables: An econometricians perspective. *Statistical Science*, 29(3):323–358, 2014.

Guido W Imbens. Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. *arXiv preprint arXiv:1907.07271*, 2019.

Guido W Imbens and Joshua D Angrist. Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475, 1994.

Guido W Imbens and Karthik Kalyanaraman. Optimal bandwidth choice for the regression discontinuity estimator. *The Review of Economic Studies*, 79 (3):933–959, 2012.

Guido W Imbens and Thomas Lemieux. Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142(2):615–635, 2008.

Guido W Imbens and Charles F Manski. Confidence intervals for partially identified parameters. *Econometrica*, 72(6):1845–1857, 2004.

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

Guido W Imbens and Stefan Wager. Optimized regression discontinuity designs. *Review of Economics and Statistics*, 101(2):264–278, 2019.

Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, 2016.

Nathan Kallus. Generalized optimal matching methods for causal inference. *arXiv preprint arXiv:1612.08321*, 2016.

Nathan Kallus and Angela Zhou. Confounding-robust policy improvement. In *Advances in Neural Information Processing Systems*, 2018.

Edward H Kennedy. Refined doubly robust estimation with undersmoothing and double cross-fitting. *preprint*, 2020.

Edward H Kennedy, Scott Lorch, and Dylan S Small. Robust causal inference with continuous instruments using the local instrumental variable curve. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(1):121–143, 2019.

Toru Kitagawa and Aleksey Tetenov. Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616, 2018.

Michal Kolesár and Christoph Rothe. Inference in regression discontinuity designs with a discrete running variable. *American Economic Review*, 108 (8):2277–2304, 2018.

Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

David S Lee and Thomas Lemieux. Regression discontinuity designs in economics. *Journal of Economic Literature*, 48(2):281–355, 2010.

Lihua Lei and Peng Ding. Regression adjustment in completely randomized experiments with a diverging number of covariates. *arXiv preprint arXiv:1806.07585*, 2018.

Fan Li, Kari Lock Morgan, and Alan M Zaslavsky. Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521):390–400, 2018.

Winston Lin. Agnostic notes on regression adjustments to experimental data: Reexamining Freedmans critique. *The Annals of Applied Statistics*, 7(1): 295–318, 2013.

Alexander R Luedtke and Antoine Chambaz. Faster rates for policy learning. *arXiv preprint arXiv:1704.06431*, 2017.

Alexander R Luedtke and Mark J van der Laan. Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Annals of Statistics*, 44(2):713, 2016.

Charles F Manski. Statistical treatment rules for heterogeneous populations. *Econometrica*, 72(4):1221–1246, 2004.

Susan A Murphy. A generalization error for Q-learning. *Journal of Machine Learning Research*, 6(Jul):1073–1097, 2005.

Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

Whitney K Newey. Efficient instrumental variables estimation of nonlinear models. *Econometrica*, 58(4):809–837, 1990.

Whitney K Newey. The asymptotic variance of semiparametric estimators. *Econometrica*, 62(6):1349–1382, 1994.

Whitney K Newey and James L Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578, 2003.

Whitney K Newey and James R Robins. Cross-fitting and fast remainder rates for semiparametric estimation. *arXiv preprint arXiv:1801.09138*, 2018.

Jersey Neyman. Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10:1–51, 1923.

Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *arXiv preprint arXiv:1712.04912*, 2017.

Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

Judea Pearl. *Causality*. Cambridge University Press, 2009.

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

Herbert Robbins. Statistical methods related to the law of the iterated logarithm. *The Annals of Mathematical Statistics*, 41(5):1397–1409, 1970.

James M Robins. A new approach to causal inference in mortality studies with a sustained exposure period: Application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9-12):1393–1512, 1986.

James M Robins. Optimal structural nested models for optimal sequential decisions. In *Proceedings of the second seattle Symposium in Biostatistics*, pages 189–326. Springer, 2004.

James M Robins and Thomas S Richardson. Alternative graphical causal models and the identification of direct effects. *Causality and Psychopathology: Finding the Determinants of Disorders and their Cures*, pages 103–158, 2010.

James M Robins and Andrea Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, 1995.

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994.

Peter M Robinson. Root-n-consistent semiparametric regression. *Econometrica*, 56(4):931–954, 1988.

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.

Paul R Rosenbaum and Donald B Rubin. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79(387):516–524, 1984.

Andrew D Roy. Some thoughts on the distribution of earnings. *Oxford Economic Papers*, 3(2):135–146, 1951.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.

Donald B Rubin. For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, 2(3):808–840, 2008.

Daniel J Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, and Zheng Wen. A tutorial on Thompson sampling. *Foundations and Trends in Machine Learning*, 11(1):1–96, 2018.

Jerome Sacks and Donald Ylvisaker. Linear estimation for approximately linear models. *The Annals of Statistics*, 6(5):1122–1137, 1978.

Daniel O Scharfstein, Andrea Rotnitzky, and James M Robins. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120, 1999.

Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, Prediction, and Search*. Springer-Verlag, New York, 1993.

Charles M Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151, 1981.

Jörg Stoye. Minimax regret treatment choice with finite samples. *Journal of Econometrics*, 151(1):70–81, 2009.

Adith Swaminathan and Thorsten Joachims. Batch learning from logged bandit feedback through counterfactual risk minimization. *The Journal of Machine Learning Research*, 16(1):1731–1755, 2015.

Donald L Thistlethwaite and Donald T Campbell. Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, 51(6):309–317, 1960.

Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 2139–2148, 2016.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

Lu Tian, Ash A Alizadeh, Andrew J Gentles, and Robert Tibshirani. A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, 109(508): 1517–1532, 2014.

Mark J van der Laan and Sherri Rose. *Targeted learning: Causal inference for observational and experimental data*. Springer Science & Business Media, 2011.

Mark J van der Laan and Daniel Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1), 2006.

Mark J van der Laan, Eric C Polley, and Alan E Hubbard. Super learner. *Statistical applications in genetics and molecular biology*, 6(1), 2007.

Stefan Wager. Cross-validation, risk estimation, and model selection: Comment on a paper by Rosset and Tibshirani. *Journal of the American Statistical Association*, 115(529):157–160, 2020a.

Stefan Wager. On regression tables for policy learning: Comment on a paper by Jiang, Song, Li and Zeng. *Statistica Sinica*, 2020b.

Stefan Wager, Wenfei Du, Jonathan Taylor, and Robert J Tibshirani. High-dimensional regression adjustments in randomized experiments. *Proceedings of the National Academy of Sciences*, 113(45):12673–12678, 2016.

Jeffrey M Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. MIT press, 2010.

Baqun Zhang, Anastasios A Tsiatis, Eric B Laber, and Marie Davidian. Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika*, 100(3):681–694, 2013.

Cun-Hui Zhang and Stephanie S Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.

Qingyuan Zhao. Covariate balancing propensity score by tailored loss functions. *The Annals of Statistics*, 47(2):965–993, 2019.

Qingyuan Zhao, Dylan S Small, and Ashkan Ertefaie. Selective inference for effect modification via the lasso. *arXiv preprint arXiv:1705.08020*, 2017.

Yingqi Zhao, Donglin Zeng, A John Rush, and Michael R Kosorok. Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499):1106–1118, 2012.

José R Zubizarreta. Using mixed integer programming for matching in an observational study of kidney failure after surgery. *Journal of the American Statistical Association*, 107(500):1360–1371, 2012.

José R Zubizarreta. Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511):910–922, 2015.