Lecture Notes on

Selection on Observables

Prof. Kyle Butts

Selection on Observables

In our study of potential outcomes, we learned about the fundamental problem of causal inference: we can never observe both $Y_i(1)$ and $Y_i(0)$ for the same unit. Randomized controlled trials solve this elegantly by making treatment assignment independent of potential outcomes, ensuring that control units provide a good estimate of what would have happened to treated units had they not been treated.

But most empirical work in economics relies on observational data where treatment is not randomly assigned. Units self-select into treatment based on their characteristics, creating systematic differences between treated and control groups that extend beyond the treatment itself. This creates selection bias: $\mathbb{E}[Y_i(0) \mid D_i = 1] \neq \mathbb{E}[Y_i(0) \mid D_i = 0]$.

Consider a concrete example: studying whether attending college increases earnings. We cannot simply compare the average earnings of college graduates to non-graduates because these groups differ in many ways beyond college attendance. College graduates likely have higher "school smarts", more educated parents, attended better high schools. These differences would likely lead to higher earnings even without the college education, i.e. higher average $Y_i(0)$.

Selection on observables offers a potential solution to this problem. The key idea is deceptively simple: if we can observe and control for all the variables that both influence treatment decisions and affect outcomes, we can eliminate selection bias. By comparing treated and control units who are on average the same in observable characteristics, we hopefully restore the conditions that make causal inference possible. However the primary concern will be whether the treated and untreated individuals are similar in terms of their *unobservable characteristics*

The Conditional Independence Assumption

Our assumption is that by comparing individuals with similar *observable* characteristics will also be similar in all their *unobservable* characteristics. This is formalized by the **conditional independence assumption (CIA)**:

$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp D_i \mid \boldsymbol{X}_i = \boldsymbol{x}$$

This says that among units with identical observable characteristics $\boldsymbol{X}_i = \boldsymbol{x}$, treatment assignment is independent of potential outcomes. In effect, we are claiming that within each "cell" defined by $\boldsymbol{X}_i = \boldsymbol{x}$, treatment is "as good as randomly assigned."

Think of this as having many small randomized experiments simultaneously – one for each value of $\boldsymbol{x}$. Among workers with the same family background, "school smarts", and high school performance, whether someone attends college might be driven by factors largely unrelated to their earnings potential. Perhaps a scholarship opportunity, a particularly inspiring teacher, living near a college, or other random family circumstances. The key assumption is that once you control for $\boldsymbol{X}$, the remaining drivers into treatment (*selection*) is due to factors that do not have an impact on earnings.

When CIA is Plausible

The credibility of CIA varies dramatically across research contexts. Some settings naturally create variation in treatment that appears unrelated to outcomes, while others involve selection processes so complex that controlling for observables seems hopeless.

Administrative processes with clear rules often provide the most credible applications. Dale and Krueger (2002) study the effect of attending elite colleges by focusing on students who applied to the same schools but made different enrollment decisions. Among students with identical college application portfolios—reflecting similar ambitions, qualifications, and family resources—the choice of which acceptance offer to take may be driven by factors largely unrelated to future earnings potential.

Institutional lotteries create natural experiments within observational data. Imbens, Rubin, and Sacerdote (2001) study how unearned income affects labor supply using lottery data. They cannot compare winners to the general population (lottery players are systematically different), but among people who played and won something, whether someone won a large or small amount may be effectively random.

Policy discontinuities exploit arbitrary cutoffs in program rules. Abdulkadiroğlu, Angrist, and Pathak (2014) study elite exam schools in New York City. These schools use test score cutoffs for admission, but the scores likely contain measurement error. Students just above and below the cutoff probably have very similar ability, making their different school assignments nearly random.

Unexpected variation can create natural experiments within complex economic processes. Anderson (2017) studies whether successful college athletics programs help universities by comparing schools that had the same expected wins (based on betting markets) but different realized outcomes. Among schools predicted to win the same number of games, those that exceeded expectations may have experienced essentially random good fortune.

These examples share a common thread: they identify sources of treatment variation that appear unrelated to potential outcomes after conditioning on observable characteristics. The key challenge is determining what variables belong in $\boldsymbol{X}_i$ and assessing whether unobserved factors remain correlated with both treatment and outcomes.

From Block Randomization to General CIA

To understand how selection on observables based estimation works in practice, it helps to start with a familiar experimental setting that illustrates the key principles.

Consider a randomized experiment where researchers vary treatment probabilities by gender: women are randomly assigned to treatment with 75% probability, while men face 25% probability. Within each gender group, treatment assignment is perfectly random and unrelated to potential outcomes. However, the overall sample lacks unconditional independence because treated and control groups have different gender compositions.

Specifically, the treated group is 75% female while the control group is only 25% female. If women and men have different baseline outcomes, this compositional difference creates bias:

$$\begin{aligned}\mathbb{E}[Y_i(0) \mid D_i = 1] &= 0.75 \cdot \mathbb{E}[Y_i(0) \mid \text{Female}] \\ &\quad + 0.25 \cdot \mathbb{E}[Y_i(0) \mid \text{Male}] \\ \mathbb{E}[Y_i(0) \mid D_i = 0] &= 0.25 \cdot \mathbb{E}[Y_i(0) \mid \text{Female}] \\ &\quad + 0.75 \cdot \mathbb{E}[Y_i(0) \mid \text{Male}]\end{aligned}$$

These are unequal whenever women and men

have different baseline outcomes, even though treatment was randomized within each group.

The solution is straightforward: estimate separate treatment effects for women and men, then aggregate using appropriate weights. For each gender group g, we compute:

$$\begin{aligned}\tau_g = \ & \mathbb{E}[Y_i \mid D_i = 1, \text{Gender} = g] \\ & - \mathbb{E}[Y_i \mid D_i = 0, \text{Gender} = g]\end{aligned}$$

These are unbiased estimates of the conditional average treatment effects (CATEs) because treatment is random within each group. To recover the overall ATE, we take a weighted average:

$$\tau_{\text{ATE}} = \tau_{\text{Female}} \cdot \mathbb{P}(\text{Female}) + \tau_{\text{Male}} \cdot \mathbb{P}(\text{Male})$$

The weights reflect each group's proportion in the overall population. For the ATT, we instead weight by the proportion of each group among treated units.

This logic extends naturally to continuous and multi-dimensional $\boldsymbol{X}_i$. For each possible value $\boldsymbol{x}$, we can estimate the conditional treatment effect by comparing treated and control units with $\boldsymbol{X}_i = \boldsymbol{x}$:

$$\begin{aligned}\tau(\boldsymbol{x}) = \ & \mathbb{E}[Y_i \mid D_i = 1, \boldsymbol{X}_i = \boldsymbol{x}] \\ & - \mathbb{E}[Y_i \mid D_i = 0, \boldsymbol{X}_i = \boldsymbol{x}].\end{aligned}$$

This is just our difference-in-means estimator for a specific value of $\boldsymbol{x}$. Under our selection on observables assumption, this equals the true CATE: $\tau(\boldsymbol{x}) = \mathbb{E}[Y_i(1) - Y_i(0) \mid \boldsymbol{X}_i = \boldsymbol{x}]$.

To recover population-level treatment effects, we integrate over the appropriate distribution:

$$\begin{aligned}\tau_{\text{ATE}} &= \int \tau(\boldsymbol{x}) \, d\mathbb{P}(\boldsymbol{X} = \boldsymbol{x}) \\ \tau_{\text{ATT}} &= \int \tau(\boldsymbol{x}) \, d\mathbb{P}(\boldsymbol{X} = \boldsymbol{x} \mid D = 1)\end{aligned}$$

The ATE uses the population distribution of $\boldsymbol{X}$, while the ATT uses the distribution among treated units. This framework shows that selection on observables methods are fundamentally about aggregating evidence from many conditional comparisons, each of which is internally valid under the CIA.

Regression Adjustment

While the stratification approach we just described works well for discrete variables like gender, it quickly becomes impractical as we add more variables or encounter continuous covariates. The **Curse of Dimensionality** strikes again‼ With many variables, most combinations of $\boldsymbol{X}_i$ values will have few or no observations, making reliable conditional comparisons impossible. This curse of dimensionality forces us to model rather than directly estimate the conditional expectation functions.

Regression adjustment operationalizes the conditional independence assumption by estimating smooth functions that predict outcomes given characteristics. Instead of stratifying the data into discrete cells, we assume the conditional expectations follow some parametric form and estimate the parameters using all available data.

The Conditional Expectation Approach

Since we have two potential outcomes, there are two distinct conditional expectations:

$$\begin{aligned}\mu_1(\boldsymbol{x}) &= \mathbb{E}[Y_i(1) \mid \boldsymbol{X}_i = \boldsymbol{x}] \\ \mu_0(\boldsymbol{x}) &= \mathbb{E}[Y_i(0) \mid \boldsymbol{X}_i = \boldsymbol{x}].\end{aligned}$$

The regression adjustment method rests on a simple insight: if we can model how outcomes de-

pend on characteristics, we can predict the missing counterfactuals. For example, we could try using the control units to estimate $\mu_0(\boldsymbol{x})$, and then predict $Y_i(0)$ for everyone. But, there's a problem. If there is selection into treatment, then the relationship between $Y_i(0)$ and $\boldsymbol{X}_i$ might differ between the control and the treated groups.

However, if we assume the CIA, then conditional on $\boldsymbol{X}_i$, outcomes are unrelated to treatment D_i. This is written as:

$$\mathbb{E}[Y_i(1)\boldsymbol{X}_i = \boldsymbol{x}, D_i = 1] = \mathbb{E}[Y_i(1)\boldsymbol{X}_i = \boldsymbol{x}]$$

$$\mathbb{E}[Y_i(0)\boldsymbol{X}_i = \boldsymbol{x}, D_i = 0] = \mathbb{E}[Y_i(0)\boldsymbol{X}_i = \boldsymbol{x}]$$

This means we can estimate each potential outcome function using only the corresponding observed data. To estimate $\mu_1(\boldsymbol{x})$, we use outcomes from treated units; to estimate $\mu_0(\boldsymbol{x})$, we use outcomes from control units.

Once we have estimates $\hat{\mu}_1(\boldsymbol{x})$ and $\hat{\mu}_0(\boldsymbol{x})$, we can predict both potential outcomes for every unit in our sample:

- For any unit i, we predict $\hat{Y}_i(1) = \hat{\mu}_1(\boldsymbol{X}_i)$ and $\hat{Y}_i(0) = \hat{\mu}_0(\boldsymbol{X}_i)$
- The predicted individual treatment effect is $\hat{\tau}_i = \hat{Y}_i(1) - \hat{Y}_i(0)$
- The ATE estimate is the average predicted effect: $\hat{\tau}_{\text{ATE}} = \frac{1}{n}\sum_{i=1}^{N} \hat{\tau}_i$
- The ATT estimate uses only treated units: $\hat{\tau}_{\text{ATT}} = \frac{1}{N_1}\sum_{D_i=1} \hat{\tau}_i$

This approach transforms the causal inference problem into a prediction problem. The quality of our treatment effect estimates depends entirely on how well we can model the conditional expectation functions.

Linear Model Implementation

The most straightforward approach assumes that the conditional expectation functions are linear in $\boldsymbol{W}_i = \boldsymbol{g}(\boldsymbol{X}_i)$. We estimate separate regressions for treated and control groups:

$$\begin{aligned} \text{Treated group:} \quad & Y_i = \alpha_1 + \boldsymbol{W}_i'\beta_1 + u_{i1} \\ \text{Control group:} \quad & Y_i = \alpha_0 + \boldsymbol{W}_i'\beta_0 + u_{i0} \end{aligned}$$

These regressions give us estimated conditional expectation functions: $\hat{\mu}_1(\boldsymbol{x}) = \hat{\alpha}_1 + \boldsymbol{w}'\hat{\beta}_1$ and $\hat{\mu}_0(\boldsymbol{w}) = \hat{\alpha}_0 + \boldsymbol{w}'\hat{\beta}_0$.

Using these fitted functions, we can compute treatment effect estimates:

$$\begin{aligned} \hat{\tau}_{\text{ATE}} &= \frac{1}{n}\sum_{i=1}^{n}[(\hat{\alpha}_1 + \boldsymbol{W}_i'\hat{\beta}_1) - (\hat{\alpha}_0 + \boldsymbol{W}_i'\hat{\beta}_0)] \\ &= (\hat{\alpha}_1 - \hat{\alpha}_0) + (\hat{\beta}_1 - \hat{\beta}_0)'\bar{\boldsymbol{W}} \end{aligned}$$

For the ATT, we focus only on treated units and rely primarily on the control group model:

$$\hat{\tau}_{\text{ATT}} = \frac{1}{n_1}\sum_{D_i=1}[Y_i - (\hat{\alpha}_0 + \boldsymbol{W}_i'\hat{\beta}_0)]$$

This takes the observed outcomes for treated units and subtracts their predicted counterfactual outcomes based on the control group regression.

Single Regression Implementation

Rather than running separate regressions, we can obtain the same estimates using a single regression with interaction terms. The key insight is to include interactions between treatment and all covariates:

$$Y_i = \alpha_0 + \alpha_1 D_i + \beta_0'\boldsymbol{W}_i + \beta_1' D_i \cdot (\boldsymbol{W}_i - \bar{\boldsymbol{W}}) + u_i$$

In this specification:

- $\hat{\alpha}_1$ directly estimates $\hat{\tau}_{\text{ATE}}$ when $\boldsymbol{W}_i$ is centered

around the population mean
- To get $\hat{\tau}_{\text{ATT}}$, center $\boldsymbol{W}_i$ around the treated group mean instead
- Standard errors are automatically correct, avoiding complex calculations needed with two-step procedures
- The coefficients have the same interpretation as the separate regression approach

This single-regression approach is typically more convenient for implementation and inference, which is why it's commonly used in practice.

Matching Methods

Regression adjustment assumes we can correctly model the conditional expectation functions, but this becomes challenging with many covariates or complex relationships. **Matching methods** offer an alternative approach that sidesteps functional form assumptions by finding "similar" control units for each treated unit based on observable characteristics.

The intuition is appealingly simple: for each person who received treatment, find someone who didn't receive treatment but looks as similar as possible on all relevant characteristics. Compare their outcomes to estimate the treatment effect. This nonparametric approach makes minimal assumptions about functional forms and can adapt to complex, unknown relationships.

The Curse of Dimensionality Problem

In principle, matching seems straightforward—just find units with identical values of $\boldsymbol{X}_i$. This exact matching works well when we have a few discrete variables, like the gender example we discussed earlier. But as we add more variables or encounter continuous covariates, exact matching becomes impossible.

Consider studying the effect of job training on earnings, controlling for age, education, work experience, family income, and region. Even if we discretize continuous variables into broad categories, we quickly generate thousands of possible combinations. Most of these combinations will contain few or no observations, making reliable comparisons impossible.

With continuous variables, the problem becomes even more severe. No two people have exactly the same age, income, and experience, so exact matches simply don't exist. We need a way to define "similarity" when exact matches are unavailable.

The matching approach solves this by finding control units with $\boldsymbol{X}_j \approx \boldsymbol{X}_i$ for each treated unit i. But this raises a new question: how do we measure similarity when characteristics are multidimensional? Different distance metrics can lead to very different matches, potentially affecting our conclusions. We will revisit the choice of distance measures later, but for now assume we have a good one.

The Matching Estimator

Once we've defined our distance metric, the matching procedure is straightforward. For each treated unit i, we find the control unit(s) with the smallest distance and use their outcome as an estimate of $Y_i(0)$.

Let $j(i)$ denote the control unit matched to treated unit i. The basic matching estimator for the ATT is:

$$\hat{\tau}_{\text{ATT}} = \frac{1}{N_1} \sum_{i=1}^{n} D_i(Y_i - Y_{j(i)})$$

where $N_1 = \sum_{i=1}^{n} D_i$ is the number of treated units.

This estimator has an elegant interpretation. For each treated unit, we observe their actual outcome

Y_i and subtract the outcome of their matched control $Y_{j(i)}$, giving an estimate of the individual treatment effect. We then average these individual estimates across all treated units.

The key insight is that matching creates a control group with the same distribution of $\boldsymbol{X}$ as the treated group *by construction*. Since we match each treated unit to controls with similar characteristics, the matched control group necessarily has characteristics similar to the treated group. This removes selection bias on observable characteristics.

Several implementation choices affect the estimator:

1. **Number of matches**: We can match each treated unit to multiple controls and average their outcomes. Using more matches reduces variance (more data) but may increase bias (worse match quality). Common choices include 1-to-1 matching, 1-to-3 matching, or 1-to-5 matching.
2. **Matching with replacement**: Should we allow the same control unit to be matched to multiple treated units? Matching with replacement ensures we always find the best possible match for each treated unit but reduces the effective sample size. Matching without replacement uses each control only once but may force some treated units to accept lower-quality matches.
3. **Caliper restrictions**: We can impose maximum distance thresholds, refusing to match if no control unit is sufficiently close. This improves match quality but may force us to discard some treated units, changing the population of interest.

Matching is a two-step procedure and that complicates standard inference: we first construct a matched sample using the covariates (or estimated distances), and then compute a difference-in-means (or run a regression) on the matched sample. Because the matching step is estimated, standard regression standard errors that treat the matches as fixed typically understate sampling variability.

Overlap and Quality Concerns

Ensuring Matches Exist:

The overlap condition requires that for every $\boldsymbol{x}$ we have $\epsilon < \mathbb{P}(D = 1 \mid \boldsymbol{X} = \boldsymbol{x}) < 1 - \epsilon$, which ensures that in large samples both treated and control observations are available for each value of $\boldsymbol{x}$. If overlap is poor, a common practical alternative is to restrict the analysis to those treated units that have good matches; doing so alters the estimand. In practice, how to actually do this is a bit ad-hoc and different paperse recommend different approaches.

Assessing Match Quality:

After matching, evaluate balance on the covariates used to form matches and also inspect balance on variables that were not included in the matching algorithm. A useful summary measure is the standardized difference

$$\frac{\bar{X}_1 - \bar{X}_0}{\sqrt{S_1^2 + S_0^2}},$$

but exercise caution with standard t-tests because, in large samples, they can be misleading.

Distance Metrics

To implement matching, we need a way to measure how "close" two units are in terms of their characteristics. The choice of distance metric can significantly affect which units get matched together, so understanding the properties of different metrics is crucial.

Euclidean Distance is the most intuitive mea-

sure:

$$d_{ij} = \sqrt{\sum_{k=1}^{K}(X_{i,k} - X_{j,k})^2} = \sqrt{(\boldsymbol{X}_i - \boldsymbol{X}_j)'(\boldsymbol{X}_i - \boldsymbol{X}_j)}$$

This is simply the straight-line distance between two points in K-dimensional space. While conceptually appealing, Euclidean distance has several serious problems in practice:

- **Not scale-invariant**: A variable measured in thousands (income) will dominate one measured in units (years of education), even if both are equally important for matching.
- **Treats all variables equally**: It gives the same weight to a one-unit difference in every variable, regardless of how much variation that variable typically exhibits.
- **Double-counts correlated variables**: If we include both "years of education" and "has college degree," highly correlated differences get counted twice in the distance calculation.

Consider an example with income (ranging 0-100,000) and education (ranging 0-20 years). A \$10,000 income difference contributes 100,000,000 to the squared distance, while a 10-year education difference contributes only 100. The income difference completely dominates, even though both might be equally important for outcomes.

Mahalanobis Distance addresses these problems by standardizing variables and accounting for correlations:

$$d_{ij} = \sqrt{(\boldsymbol{X}_i - \boldsymbol{X}_j)' S_{\boldsymbol{XX}}^{-1} (\boldsymbol{X}_i - \boldsymbol{X}_j)}$$

where $S_{\boldsymbol{XX}}$ is the sample variance-covariance matrix of $\boldsymbol{X}$. The Mahalanobis distance essentially transforms the covariate space so that all variables have unit variance and are uncorrelated. This transformation makes the distance measure scale-invariant and does not suffer "double counting" problem. In practice, Mahalanobis distance typically produces better matches and more reliable treatment effect estimates, which is why it's become the standard choice for matching applications.

Propensity Score Methods

Matching on the full covariate vector $\boldsymbol{X}_i$ becomes impractical when we have many variables or complex relationships. Even with sophisticated distance metrics, finding good matches in high-dimensional spaces is challenging—a problem known as the "curse of dimensionality."

Rosenbaum and Rubin (1983) provided an elegant solution: instead of matching on the full vector $\boldsymbol{X}_i$, we can match on a single number that summarizes all relevant information about treatment assignment. This **propensity score** dramatically simplifies matching while preserving the conditional independence assumption.

The Propensity Score

The **propensity score** is defined as the probability of treatment given observed characteristics:

$$\pi(\boldsymbol{x}) = \mathbb{P}(D_i = 1 \mid \boldsymbol{X}_i = \boldsymbol{x})$$

At first glance, this seems like we're throwing away information by reducing a multi-dimensional vector to a single number. Remarkably, Rosenbaum and Rubin showed that this single number contains all the information we need for causal inference under selection on observables. Formally, if selection on observables,

$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp D_i \mid \boldsymbol{X}_i,$$

then selection on the propensity score holds too:

$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp D_i \mid \pi(\boldsymbol{X}_i).$$

This means we can achieve the same identifying power by conditioning on $\pi(\boldsymbol{X}_i)$ instead of the full vector $\boldsymbol{X}_i$. Units with the same propensity score are comparable for treatment effect estimation, regardless of their specific combination of characteristics.

Estimating Propensity Scores

In practice, we don't know the true propensity score $\pi(\boldsymbol{X}_i)$ and must estimate it from data. Since treatment is a binary outcome, logistic regression is the natural choice:

$$\pi(\boldsymbol{x}) = \frac{\exp(\boldsymbol{w}'\gamma)}{1 + \exp(\boldsymbol{w}'\gamma)},$$

where $\boldsymbol{w}$ is again a transformation of the conditioning $\boldsymbol{X}$.

This logistic specification ensures that fitted values lie between 0 and 1, as probabilities must. The key is to include sufficiently flexible terms to capture the true relationship between characteristics and treatment probability. Just as with outcome models in regression adjustment, functional form matters crucially for propensity score estimation. The quality of the propensity score model affects all subsequent analysis.

Propensity Score Matching

Once we have estimated propensity scores $\hat{\pi}(\boldsymbol{X}_i)$, matching becomes remarkably simple. Instead of matching on the high-dimensional vector $\boldsymbol{X}_i$, we match on the scalar $\hat{\pi}(\boldsymbol{X}_i)$.

The procedure is:

1. Estimate propensity scores for all units using logistic regression
2. For each treated unit, find the control unit(s) with the closest estimated propensity score
3. Compute the treatment effect as the difference in average outcomes between treated units and their matched controls

This approach has several advantages over matching on $\boldsymbol{X}_i$ directly:

- **Dimension reduction**: We match on one number instead of many variables
- **Interpretable distances**: Propensity score differences have natural interpretations as differences in treatment probability
- **Visual assessment**: We can easily plot propensity score distributions to assess overlap
- **Computational simplicity**: Finding the closest scalar is much faster than computing high-dimensional distances

The same practical considerations apply as with direct matching: we need to decide how many matches to use, whether to match with replacement, and whether to impose caliper restrictions on maximum propensity score differences.

Inverse Probability Weighting

Matching discards control units that aren't selected as matches, potentially wasting valuable information. **Inverse probability weighting (IPW)** offers an alternative approach that uses all available data by reweighting observations to eliminate selection bias.

The key insight behind IPW is surprisingly simple. In a randomized experiment, all units have the same probability of treatment, so treated and control groups are directly comparable. In observational data, different units have different treatment probabilities based on their characteristics. IPW adjusts for this by giving more weight to units whose treatment status is "surprising" given their characteris-

tics.

Consider a control unit with characteristics that make treatment very likely—say, a high-ability student from a wealthy family applying to college. If this person doesn't attend college despite high probability, they provide especially valuable information about what would happen to treated units if they weren't treated. IPW gives such observations higher weight in our analysis.

The IPW estimator for the ATE is:

$$\hat{\tau}_{\text{IPTW}} = \mathbb{E}_n\left[\frac{D_i Y_i}{\pi(\boldsymbol{X}_i)}\right] - \mathbb{E}_n\left[\frac{(1-D_i)Y_i}{1-\pi(\boldsymbol{X}_i)}\right]$$

The first term reweights treated units by $1/\pi(\boldsymbol{X}_i)$, giving more weight to treated units with low treatment probability. The second term reweights control units by $1/(1-\pi(\boldsymbol{X}_i))$, giving more weight to control units with high treatment probability.

For the ATT, we modify the control group weighting to match the treated group distribution:

$$\hat{\tau}_{\text{ATT}} = \mathbb{E}_n\left[\frac{D_i Y_i}{n_1}\right] - \mathbb{E}_n\left[\frac{\pi(\boldsymbol{X}_i)(1-D_i)Y_i}{(1-\pi(\boldsymbol{X}_i))}\right]$$

Here, we weight control units by $\pi(\boldsymbol{X}_i)/(1-\pi(\boldsymbol{X}_i))$, which upweights controls that look more like treated units.

Practical Concerns

IPW methods work well when propensity scores are moderate, but they can become problematic when some units have very high or very low treatment probabilities.

Extreme Weights: When $\pi(\boldsymbol{X}_i) \approx 0$ or $\pi(\boldsymbol{X}_i) \approx 1$, the weights $1/\pi(\boldsymbol{X}_i)$ or $1/(1-\pi(\boldsymbol{X}_i))$ become extremely large. A unit with $\pi(\boldsymbol{X}_i) = 0.001$ receives a weight of 1000, potentially dominating the entire analysis. This creates very noisy estimates where a few observations with extreme weights drive all the results.

Common solutions include:

- **Trimming**: Drop observations with propensity scores below some threshold (e.g., 0.1) or above some threshold (e.g., 0.9)
- **Winsorizing**: Cap extreme weights at some maximum value
- **Hájek weights**: Normalize weights to sum to one, which reduces but doesn't eliminate the extreme weight problem

Overlap Assessment: Before implementing IPW, it's crucial to examine whether treated and control groups have overlapping propensity score distributions. Plot histograms or density plots of estimated propensity scores separately for treated and control groups. If there's little overlap—for instance, if all treated units have high propensity scores while all controls have low scores—then IPW will rely heavily on extrapolation and produce unreliable results.

Why Not Linear Regression?

Given all the "complexity" of matching, propensity scores, and regression adjustment, a natural question arises: if we believe the conditional independence assumption holds, why not simply run a basic linear regression?

$$Y_i = \tau D_i + \boldsymbol{W}_i'\beta + u_i$$

This approach is appealingly simple and widely used in practice. Is $\hat{\tau}$ a reasonable estimate of the ATE or the ATT? Understanding what this regression actually estimates and comparing it to the methods we've discussed, reveals important limitations that explain why specialized selection-on-observables methods are often preferable.

What Does OLS Estimate?

The Frisch-Waugh-Lovell theorem tells us exactly what the OLS coefficient on treatment identifies. The estimated treatment effect can be written as:

$$\hat{\tau}_{\text{OLS}} = \mathbb{E}\left[\frac{\tilde{D}_i}{\text{Var}(\tilde{D}_i)} Y_i\right]$$

where $\tilde{D}_i$ is the residual from regressing D_i on $\boldsymbol{X}_i$. This is a weighted average of outcomes where the weights depend on how "surprising" each unit's treatment status is given their characteristics.

Units with large values of $|\tilde{D}_i|$, i.e. those whose treatment status is poorly predicted by their observable characteristics, receive disproportionate weight in determining the treatment effect estimate. A treated unit with characteristics that make treatment unlikely, or a control unit with characteristics that make treatment likely, will heavily influence our results.

This weighting scheme has important implications that differ from the methods we've studied.

Homogeneous vs. Heterogeneous Effects

Whether OLS provides a good estimator depends crucially on whether treatment effects are constant across units.

Under homogeneous treatment effects, where $\tau_i = \tau$ for all units, the weighting scheme doesn't matter. Every unit has the same treatment effect, so any weighted average of treatment effects equals the true ATE. In this case, $\hat{\tau}_{\text{OLS}} \to^p \tau_{\text{ATE}}$ and linear regression provides a consistent, efficient estimator.

Under heterogeneous treatment effects, however, OLS estimates a complex weighted average:

$$\xrightarrow{p} \hat{\tau}_{\text{OLS}} = \mathbb{E}[w_i \tau_i] / \mathbb{E}[w_i]$$

where the weights $w_i = (D_i - \mathbb{E}[D_i \mid \boldsymbol{X}_i])^2$ depend on prediction errors from the propensity score.

This weighted average is generally *not* the ATE, ATT, or any other standard policy parameter. The weights are determined by the stochastic process generating treatment assignment rather than policy considerations. Units whose treatment decisions are hardest to predict get the most influence, regardless of whether their experiences are representative or policy-relevant.

The Aronow and Samii Critique

Aronow and Samii (2016) analyzed linear regression under the most favorable possible conditions for the method. They assumed:

- The conditional independence assumption holds perfectly
- Untreated outcomes are modelled correction: $Y_i(0) = \boldsymbol{X}_i'\beta + u_i$
- Treatment probabilities are linear in covariates: $\pi(\boldsymbol{X}_i) = \boldsymbol{X}_i'\gamma$

Even under these ideal conditions, the OLS estimator fails to identify policy-relevant parameters when treatment effects are heterogeneous.

Their empirical illustration using cross-country growth regressions is particularly striking. They show that OLS estimates can be dominated by a few countries with unusual combinations of characteristics and treatment status. Small island nations or oil-rich economies with atypical institutions might drive conclusions about the relationship between democracy and growth, even though their experiences may not generalize to the countries policymakers actually care about.

Note that none of the the previous estimators create this "weirdly weighted" average of treated units. In practice, these other estimators are preferreable (at least in my opinion), but you will still see regression a lot in practice.