

MY457/MY557: Causal Inference for Observational and Experimental Studies

Week 4: Selection on Observables 2

Daniel de Kadt

Department of Methodology
LSE

Winter Term 2024

Topics of this lecture

- 1 Where Were We?
- 2 Matching
- 3 Weighting
- 4 Regression
- 5 Practical Advice

Where Were We?

Observational settings where the **assignment mechanism** for D is either **unknown** or **not under our control**.

Problem: If Y_1 , Y_0 , and D are associated with **observed pre-treatment** X (a 'selection problem'), we **cannot naively compare** the group means of Y .

Solution: We make the (1) **conditional independence assumption**:

$$(Y_1, Y_0) \perp\!\!\!\perp D \mid X$$

And (2) **common support** assumption:

$$0 < \Pr(D_i = 1 \mid X_i = x) < 1 \quad \text{for any } x \in \mathcal{X}$$

But even then, we still **cannot** naively compare the means of Y in different groups!

Where Were We?

Given our assumptions, the ATE is instead **nonparametrically identified** as the weighted difference in **population regression functions**:

$$\begin{aligned}\tau_{ATE} &= \mathbb{E}[\hat{\tau}_{CATE}(X_i)] \\ &= \int (\mathbb{E}[Y_i \mid D_i = 1, X_i = x] - \mathbb{E}[Y_i \mid D_i = 0, X_i = x])f(x)dx\end{aligned}$$

The intuition easier to grasp if we consider a case in which all X_i is **discrete**...

Then we can **rewrite** the identification result (for both ATE and ATT) as:

$$\begin{aligned}\tau_{ATE} &= \sum_{x \in \mathcal{X}} (\mathbb{E}[Y_i \mid D_i = 1, X_i = x] - \mathbb{E}[Y_i \mid D_i = 0, X_i = x]) \Pr(X_i = x) \\ \tau_{ATT} &= \sum_{x \in \mathcal{X}} (\mathbb{E}[Y_i \mid D_i = 1, X_i = x] - \mathbb{E}[Y_i \mid D_i = 0, X_i = x]) \Pr(X_i = x \mid D_i = 1)\end{aligned}$$

We calculate conditional ATEs (CATEs) for different levels of X , and re-weight them by the (conditional) prevalence of X in the data.

From Identification to Estimation

This gives us the **identification result** for selection on observables. But how should we **estimate** our estimands of interest?

There are four broad approaches for **estimation under conditioning**:

1. Subclassification
2. Matching
3. Weighting
4. Regression

Subclassification only works with discrete X variables, and is a sample analogue of the result we saw on the previous slide (consult last week's slides).

More general, and more frequently encountered solutions are **matching**, **weighting**, and **regression**. We turn to those now.

1 Where Were We?

2 Matching

3 Weighting

4 Regression

5 Practical Advice

Matching

Matching seeks to **impute missing potential outcomes** using the observed outcomes of 'closest' units or **nearest neighbors**. Basic process:

1. For each observation in the treated group i , find an observation in the untreated group with the **most similar** values of X
- 2a. Estimate ATT with the average difference between the pairs:

$$\hat{\tau}_{ATT} = \frac{1}{n_1} \sum_{i:D_i=1} (Y_i - \tilde{Y}_i) \simeq \frac{1}{n_1} \sum_{i:D_i=1} (Y_{1i} - Y_{0i}) = \tau_{ATT}$$

where \tilde{Y}_i is the observed outcome of i 's untreated 'buddy'

- 2b. When there are multiple (M_i) 'close' units, their average can be used:

$$\hat{\tau}_{ATT} = \frac{1}{n_1} \sum_{i:D_i=1} \left(Y_i - \left(\frac{1}{M_i} \sum_{m=1}^{M_i} \tilde{Y}_{i_m} \right) \right)$$

where \tilde{Y}_{i_m} is i 's m th untreated buddy

Example with Single Pre-treatment Covariate

unit	Potential Outcome under Treatment	Potential Outcome under Control		
i	$Y_i(1)$	$Y_i(0)$	D_i	X_i
1	6	?	1	3
2	1	?	1	1
3	0	?	1	4
4		0	0	2
5		9	0	3
6		1	0	-2
7		1	0	-4

Match and plug in:

$$\hat{\tau}_{ATT} = \frac{1}{3}((6 - 9) + (1 - 0) + (0 - 9)) = -3.7$$

A Silver Bullet?

Matching looks like it is magic, but it's not.

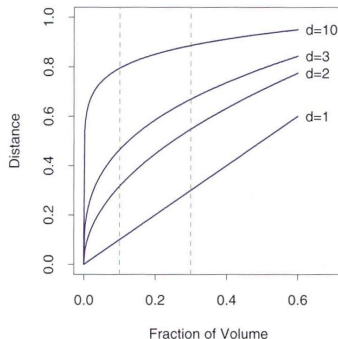
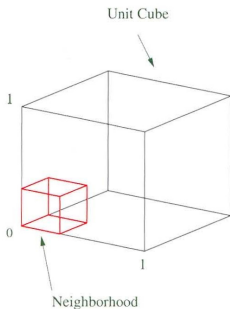
Matching is an approach to estimation (just like regression).

Always remember: "Design precedes estimation."

The Curse of Dimensionality

Consider a case where \mathbf{X}_i contains > 1 variable? Can we hope to **exactly match** on every \mathbf{X}_i , even if we have very large n ? **No!**

We are struck by what is called the **curse of dimensionality**...



As number of dimensions in the covariate space increases, data sparsity **exponentially increases** for a given sample size.



"You must prepare to settle for a 60-70% match"

The Curse of Dimensionality and Bias

The curse of dimensionality implies a **bias** problem wherever we allow for non-exact matches.

Why? By tolerating **not-quite-exact matches**, we must (in expectation) inject 'error' into our estimates of missing potential outcomes (Abadie & Imbens, 2006).

The bias term is order $N^{(-1/k)}$, **increasing in the number of dimensions k** and implying no \sqrt{n} -consistency for $k > 2$.

If N_0 is much larger than N_1 (and there is common support), bias will typically be small. **Generally wise** to use Abadie & Imbens (2011) **bias correction** (more later).

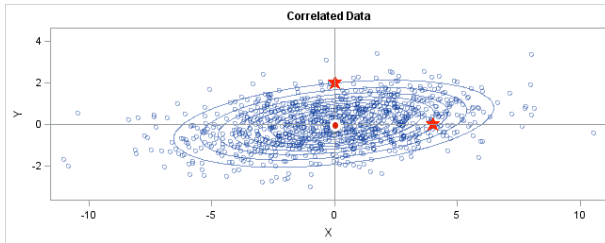
Matching as Dimension Reduction

How do we find the 'closest' match in **multi-dimensional** space?

We typically use a low-dimensional representation or **distance metric**. One example is **Mahalanobis distance**:

$$D_M(X_i, X_j) = \sqrt{(X_i - X_j)^\top \Sigma_X^{-1} (X_i - X_j)}$$

where Σ_X is the (sample) variance-covariance matrix of X_i



Note: other variants and metrics are possible.

The Propensity Score and the Balancing Property

Definition (propensity score)

Probability of receiving the treatment given X_i

$$\pi(X_i) \equiv \Pr(D_i = 1 \mid X_i)$$

Assumptions: Suppose the following holds:

1. $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i \mid X_i$ (conditional ignorability)
2. $0 < \Pr(D_i = 1 \mid X_i = x) < 1$ for any x (common support)

Result: The propensity score has the **balancing property** (Rosenbaum & Rubin, 1983):

$$D_i \perp\!\!\!\perp X_i \mid \pi(X_i)$$

Read: Among those units with the same propensity score, X_i is independent of treatment assignment.

Identification with the Propensity Score

The balancing property implies that **conditional ignorability** holds, conditional on just the **propensity score** alone:

$$(Y_{1i}, Y_{0i}) \perp\!\!\!\perp D_i \mid \pi(X_i)$$

Implication: It is **sufficient to condition on $\pi(X_i)$** , instead of X_i

But there is a catch: $\pi(X_i)$ itself needs to be **estimated**!

Two-step procedure to estimate causal estimands:

- (1) Estimate $\pi(X_i)$ with a model for a binary response (e.g. logit, probit)
- (2) Do nearest neighbor matching on $\pi(X_i)$

Note: Need to allow some uncertainty from (1) to percolate through to (2) (this is an open area of study)

Estimating the Propensity Score

Estimation of propensity scores requires a correct specification of $\pi(X_i)$ (functional form, etc.).

Check balance:

- Ideally, want to compare the joint distribution of all X_i between the treated and untreated in the matched sample
- In practice, check various low-dimensional summaries of $F(x)$ (mean difference, variance ratio, etc.)
- **Balance tests** are often used (e.g. t-test, F-test, KS test) like the ones we saw for randomized experiments.
- Note that balance tests can be misleading in a matching context because “balance” often improves when you drop lots of observations – can think of this as a “balance-sample size frontier” (King et al., 2017)

Estimate \rightarrow Check Balance \rightarrow Re-estimate \rightarrow Check Balance $\rightarrow \dots$ (*ad infinitum* until you attain good balance)

Is this data snooping or p-hacking? No, as long as inference remains blind to Y and τ

Things to Consider as you Match

There is a plethora of choices to be made:

- One-to-one vs. Many-to-one matching
 - Exact matching vs. non-exact matching
 - Matching with or without replacement
 - Caliper matching
 - Propensity score matching
 - Genetic matching
 - Optimal matching
 - Coarsened exact matching
- ...and more in the pipeline...

This creates many **researcher degrees of freedom**. Whatever you choose, do so for principled reasons (e.g. balance) and without 'snooping' (looking at $\hat{\tau}$).

Balance testing when matching is important, but can be misleading. If you only check **things you matched on** you will often see good balance. But what are you missing?

Consider the **balance-sample size frontier**: one way to achieve good balance is to heavily trim your sample. Is this a good idea? (e.g. King, Lucas, & Nielsen 2017)

1 Where Were We?

2 Matching

3 Weighting

4 Regression

5 Practical Advice

Weighting on the Propensity Score

So far we have used the propensity score for matching.

An alternative approach is **weighting**

Result: Under the conditional ignorability and common support assumptions, we can identify the ATE and ATT (weakly assuming) as:

$$\begin{aligned}\tau_{ATE} &= \mathbb{E} \left[Y_i \cdot \frac{D_i - \pi(X_i)}{\pi(X_i) \cdot (1 - \pi(X_i))} \right] \\ \tau_{ATT} &= \frac{1}{\Pr(D = 1)} \cdot \mathbb{E} \left[Y_i \cdot \frac{D_i - \pi(X_i)}{1 - \pi(X_i)} \right]\end{aligned}$$

These can be estimated using sample analogues called **inverse probability weighting (IPW)** estimators:

$$\begin{aligned}\hat{\tau}_{ATE} &= \frac{1}{N} \sum_{i=1}^N \left(Y_i \cdot \frac{D_i - \hat{\pi}(X_i)}{\hat{\pi}(X_i) \cdot (1 - \hat{\pi}(X_i))} \right) = \frac{1}{N} \sum_{i=1}^N \left(\frac{D_i Y_i}{\hat{\pi}(X_i)} - \frac{(1 - D_i) Y_i}{1 - \hat{\pi}(X_i)} \right) \\ \hat{\tau}_{ATT} &= \frac{1}{N_1} \sum_{i=1}^N \left(Y_i \cdot \frac{D_i - \hat{\pi}(X_i)}{1 - \hat{\pi}(X_i)} \right) = \frac{1}{N_1} \sum_{i=1}^N \left(D_i Y_i - (1 - D_i) Y_i \frac{\hat{\pi}(X_i)}{1 - \hat{\pi}(X_i)} \right)\end{aligned}$$

Performance of the IPW estimators

IPW estimators have poor **small sample properties**:

- They **highly sensitive to extreme values** of $\pi(X_i)$
- Tends to occur when there is a **lack of overlap**
- This generates high **variance** (inefficiency)
- Can also produce significant **bias** in certain settings (e.g. model misspecification)

A workaround is trim units with extreme weights. But this changes the estimand to a quantity that is still causal yet difficult to interpret.

Alternative weighting methods with preferable finite sample properties include:

- Augmented IPW estimators: e.g. doubly robust estimator (more later).
- Entropy balancing (Hainmueller 2012, **eбал**): choose weights that directly optimize balance in X_i .
- Covariate balancing propensity scores (Imai and Ratkovic 2014, **CBPS**): model $\pi(X_i)$ while optimizing balance in X_i .
- Kernel balancing (Hazlett, 2020, **kбал**): choose weights to balance an unspecified non-linear representation of X_i .

- 1 Where Were We?
- 2 Matching
- 3 Weighting
- 4 Regression
- 5 Practical Advice

Model-based Estimation of Causal Effects

When we think of ‘controlling for’ variables, we usually think of **regression**. What role can it play in causal inference?

Recall that under conditional ignorability and common support, ATE/ATT equal weighted averages of the differences in **population regression functions**:

$$\hat{\tau}(x) = \mathbb{E}[Y_i \mid D_i = 1, X_i = x] - \mathbb{E}[Y_i \mid D_i = 0, X_i = x]$$

where

$$\tau_{ATE} = \mathbb{E}[\hat{\tau}(x)] \quad \text{and} \quad \tau_{ATT} = \mathbb{E}[\hat{\tau}(x) \mid D_i = 1]$$

This suggests a **model-based** approach for estimating causal effects, where we use a regression model for $\mathbb{E}[Y_i \mid D_i, X_i]$, e.g.,

$$\mathbb{E}[Y_i \mid D_i, X_i] = \beta_0 + \beta_1 D_i + \mathbf{X}_i \gamma,$$

which is a linear regression, and we can estimate β_1 via OLS.

OLS as an Estimator of Causal Effects

Suppose we regressed Y_i on D_i and X_i , estimating the coefficient on D_i via OLS:

$$\hat{\beta}_{OLS} = \frac{\text{Cov}(Y_i, \tilde{D}_i)}{\text{Var}(\tilde{D}_i)},$$

where \tilde{D}_i is the residual from the regression of D_i on X_i (“partialling out”).

When is $\hat{\beta}_{OLS}$ a good estimator of τ_{ATE} ?

The answer depends on whether these two assumptions hold:

- (1) **Constant treatment effect**: $\tau = Y_{1i} - Y_{0i}$ for all i .
- (2) **Linearity**: Potential outcomes can be written as

$$Y_i(d) = \beta_0 + d\beta_1 + \mathbf{X}_i\gamma + \varepsilon_i \quad \text{for } d = 0, 1.$$

Noting that (2) implies (1) (such that $\beta_1 = \tau$), there are 3 possible scenarios:

- 1 Both (1) and (2) are true.
- 2 Only (1) is true.
- 3 Neither (1) nor (2) is true.

Case 1: Constant Effect & Linear Potential Outcomes

Result: If treatment effect is constant across units and potential outcomes are linear in \mathbf{X}_i , then the OLS estimate of β_1 in the following regression model

$$Y_i = \beta_0 + \beta_1 D_i + \mathbf{X}_i \gamma + \varepsilon_i$$

is an **unbiased and consistent** estimator of τ_{ATE} .

Proof: First, note that $\beta_1 = \tau_i$ for every i under these assumptions:

$$\begin{aligned}\tau_i &= Y_{1i} - Y_{0i} \\ &= (\beta_0 + \beta_1 + \mathbf{X}_i \gamma + \varepsilon_i) - (\beta_0 + \mathbf{X}_i \gamma + \varepsilon_i) \\ &= \beta_1\end{aligned}$$

Next, note that conditional ignorability implies the conditional independence between D_i and ε_i :

$$(Y_{1i}, Y_{0i}) \perp\!\!\!\perp D_i \mid \mathbf{X}_i \implies \varepsilon_i \perp\!\!\!\perp D_i \mid \mathbf{X}_i$$

Because this implies the zero conditional mean assumption, $\hat{\beta}_{OLS}$ is an unbiased and consistent estimator of β_1 , which is equal to τ_{ATE} (and τ_i). □

Case 2: Constant Effect & Unknown Functional Form

What happens if $Y_i(d)$ is an **unknown, nonlinear function** of d and X_i , and yet we used $\hat{\beta}_{OLS}$ as an estimator of $\hat{\tau}_{ATE}$ anyway?

Recall that OLS is the **best linear predictor** in terms of MSE:

$$\hat{\beta}_{OLS} = \underset{\hat{\beta}_1}{\operatorname{argmin}} \mathbb{E}[(Y_i - \hat{\beta}_0 - \hat{\beta}_1 D_i - \mathbf{X}_i \hat{\gamma})^2]$$

This, it turns out, also implies that $\hat{\beta}_{OLS}$ provides the **best linear approximation** to the population regression function:

$$\hat{\beta}_{OLS} = \underset{\hat{\beta}_1}{\operatorname{argmin}} \mathbb{E}[(\mathbb{E}[Y_i | D_i, X_i] - \hat{\beta}_0 - \hat{\beta}_1 D_i - \mathbf{X}_i \hat{\gamma})^2]$$

Result:

- $\hat{\beta}_{OLS}$ can be interpreted as the best linear approximation to the true treatment effect, whatever the true functional form is.
- This approximation may or may not be good in absolute terms.
- More flexible models (nonlinear, semi-/non-parametric, etc.) may provide a better performing approximation.

Case 3: Heterogeneous Treatment Effects

Consider again our default OLS specification:

$$Y_i = \beta_0 + \beta_1 D_i + \mathbf{X}_i \gamma + \varepsilon_i$$

This can be thought of as a **parametric model** of the underlying **data generating process** that produces Y_i (and by implication, Y_{1i} and Y_{0i}).

By modeling the relationship between D_i and Y_i as a multiplicative function of just β_1 , we assert that the effect of D_i is **fixed and homogeneous**.

Treatment effect heterogeneity is any real deviation from that assumed model, for example:

- 1 SUTVA violations generate variation in treatment effects
- 2 Effects vary across individual by chance
- 3 Effects vary over time (e.g. early vs. late)
- 4 Effects vary systematically by covariates (observed or unobserved)

Case 3: Heterogeneous Treatment Effects

Now recall the subclassification estimator for the **ATE**:

$$\hat{\tau}_{ATE} = \sum_{x \in \mathcal{X}} (\mathbb{E}[Y_i | D_i = 1, X_i = x] - \mathbb{E}[Y_i | D_i = 0, X_i = x]) \Pr(X_i = x),$$

where we weighted subgroup effects by the **marginal of X_i** .

Similarly, the subclassification estimator for the **ATT**:

$$\hat{\tau}_{ATT} = \sum_{x \in \mathcal{X}} (\mathbb{E}[Y_i | D_i = 1, X_i = x] - \mathbb{E}[Y_i | D_i = 0, X_i = x]) \Pr(X_i = x | D_i = 1),$$

where we weighted subgroup effects by the **conditional of X_i given $D_i = 1$** .

Result: The OLS estimator can be written as a subclassification estimator, weighted by the **conditional variances of D_i** in each subgroup (Angrist, 1998):

$$\hat{\beta}_{OLS} = \sum_{x \in \mathcal{X}} (\mathbb{E}[Y_i | D_i = 1, X_i = x] - \mathbb{E}[Y_i | D_i = 0, X_i = x]) \frac{\text{Var}(D_i | X_i = x) \Pr(X_i = x)}{\sum_{x'} \text{Var}(D_i | X_i = x') \Pr(X_i = x')}$$

OLS as a Subclassification Estimator

Estimator	Weights for Subgroups	Unbiased for
$\hat{\tau}_{ATE}$	$\Pr(X_i = x)$	τ_{ATE}
$\hat{\tau}_{ATT}$	$\Pr(X_i = x \mid D_i = 1)$	τ_{ATT}
$\hat{\beta}_{OLS}$	$\frac{\text{Var}(D_i \mid X_i = x) \Pr(X_i = x)}{\sum_{x'} \text{Var}(D_i \mid X_i = x') \Pr(X_i = x')}$	" $\tau_{CVW-ATE}$ "

With non-constant treatment effects, OLS provides an unbiased estimator for a **conditional-variance-weighted average treatment effect**.

This is a causal quantity, but hard to interpret. It is not generally equal to the ATT or ATE (more in a moment).

Recall $\text{Var}(D_i \mid X_i = x) = \pi(x)(1 - \pi(x))$. Therefore:

- Weights are high for groups with propensity scores close to **0.5**.
- Weights are low for groups with propensity scores close to **0** or **1**.
- OLS minimizes estimation uncertainty by downweighting groups where group-specific ATEs are less precisely estimated.

This result assumes discrete **X**s, but intuition holds for continuous **X**s.

OLS as a Weighted Average of Estimands

Given heterogeneous treatment effects (and some linearity assumptions), the causal estimand targeted by OLS can be decomposed as:

$$\tau_{OLS} = w_1 \cdot \tau_{ATT} + w_0 \cdot \tau_{ATU}$$

where:

$$w_1 = \frac{(1-P(D=1)) \cdot \text{Var}[\pi(X)|D=0]}{P(D=1) \cdot \text{Var}[\pi(X)|D=1] + (1-P(D=1)) \cdot \text{Var}[\pi(X)|D=0]}, \text{ and}$$
$$w_0 = 1 - w_1$$

With heterogeneous treatment effects, OLS can be an unbiased estimator for a **weighted average of the ATT and ATU** (Słoczyński, 2022).

This can admit a strange interpretation:

- Weights w_j are inversely proportional to the share of units in j .
- This is **weird**: if you have a lot of treated units, ATU will be upweighted, and ATT downweighted. Why?
- For the ATT, OLS is predicting the *missing potential outcomes* for the treated – those come from the coefficients for the control, so these are upweighted.

Solutions: weighting, matching, and fully interacting de-means \mathbf{X} and \mathbf{D} .

The Fully-Interacted Estimator

One well regarded large-sample linear regression estimator is as follows:

$$Y_i = \hat{\alpha} + \hat{\tau}_{int} D_i + \hat{\beta}^T (X_i - \bar{X}) + \hat{\gamma}^T D_i (X_i - \bar{X})$$

where:

X_i are covariates sufficient to satisfy the conditional independence assumption

\bar{X} is the sample mean of X_i

This estimator has numerous desirable properties:

- The bias in $\hat{\tau}_{int}$ as an estimator for τ_{ATE} is arbitrarily small in large samples under only conditional independence.
- Huber-White robust standard errors are sufficient for hypothesis testing.
- Mitigates small sample biases and inefficiency (Freedman, 2008).
- Resolves the weighted average of estimands problem (Słoczyński, 2022).
- Robust to contamination bias (Goldsmith-Pinkham et al, 2022)

- 1 Where Were We?
- 2 Matching
- 3 Weighting
- 4 Regression
- 5 Practical Advice

Matching or Regression?

Regression:

- + Regression is simple.
- In SOO world, simple regression relies on a number of strong assumptions to admit a readily interpretable estimate. More complex specifications can help.
- Regression is prone to extrapolation beyond common support.

Matching:

- + Non-parametric (no model dependence)
- + Can be a transparent way to move from data/design to an estimate
- Can be rather non-transparent if implemented in certain ways
- Recall that because we can very rarely ever exactly match, matching usually induces **bias** by pulling our estimate slightly away from the estimand. This becomes more severe:
 - the more matches for each treated unit (as in, $M=2$ or 3 or 10); and
 - the more covariates we match on

Combining Regression, Matching and Weighting

Some approaches combine regression with matching or weighting for better finite-sample performance and/or robustness properties.

- **Bias-corrected** matching (Abadie and Imbens 2005):
 - Estimate bias inherent to matching estimators via regression
 - Subtract it off from the matching estimate for correction
 - In **R**, can e.g. use `BiasAdjust = TRUE` in the **Matching** package.
- **Doubly-robust** estimation (Robins and Rotnitzky 2001):
 - Use a weighted average of regression and IPW estimators
 - The estimator will be consistent as long as either the regression model or PS model is correct
 - In **R**, see e.g. `tmle` or `drgee` packages.
- Matching as nonparametric data **preprocessing** (Ho, Imai, King, & Stuart 2007):
 - Model-based estimation of causal effect is most likely to go wrong when it involves **extrapolation** due to poor overlap in covariates
 - Use matching to make treatment and control groups similar
 - Then run regression models to estimate causal effects
 - In **R**, use whatever matching tool then whatever parametric tool!