Political Methodology II
Section: Selection on Observables

Apoorva Lal

January 26, 2022

Stanford University

Roadmap

Overview of Selection on Observables

Identification Assumptions

Subclassification and Matching

Propensity Scores

Regression

Roadmap

Overview of Selection on Observables

Identification Assumptions

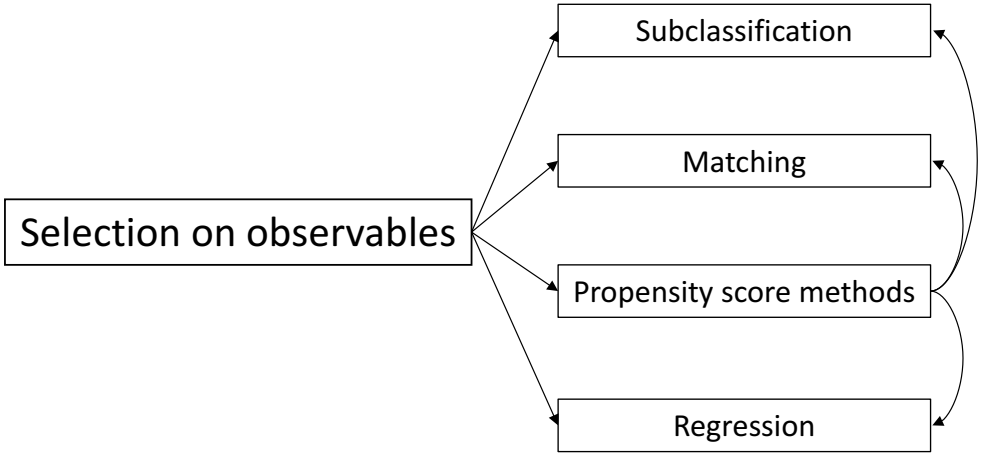
Subclassification and Matching

Propensity Scores

Regression

Idea of Selection on Observables

- Causal inference is all about understanding the treatment assignment mechanism.
- Selection on observables says that once we condition on some observable covariates, treatment assignment was as good as random.
- Requires substantive justification – if there's selection on unobservables (particularly the potential outcomes), matching won't help.
- Estimation methods include subclassification, matching, propensity score methods, and regression.


Selection on observables
Subclassification
Matching
Propensity score methods
Regression

Commonalities and Differences

Commonalities among these estimation strategies:

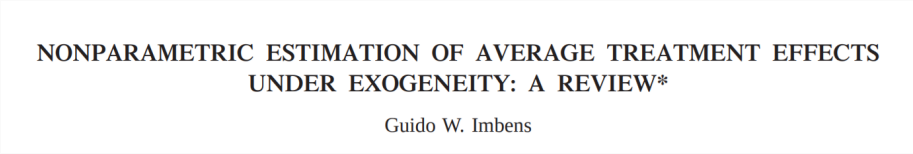
- They're ways of *imputing* the counterfactual potential outcome for treatment units by adjusting for covariates.
- They all impute this counterfactual by using a *weighted average* of observed outcomes for other units.
- Matching and subclassification are nonparametric estimators of the counterfactual, while linear regression is a parametric / semi-parametric estimator.

Commonalities and Differences

Commonalities among these estimation strategies:

- They're ways of *imputing* the counterfactual potential outcome for treatment units by adjusting for covariates.
- They all impute this counterfactual by using a *weighted average* of observed outcomes for other units.
- Matching and subclassification are nonparametric estimators of the counterfactual, while linear regression is a parametric / semi-parametric estimator.

The main difference is that they use different methods of picking the weights – making different assumptions about the functional form relating the covariates, treatment, and outcomes. Review paper from syllabus


NONPARAMETRIC ESTIMATION OF AVERAGE TREATMENT EFFECTS UNDER EXOGENEITY: A REVIEW*

Guido W. Imbens

Roadmap

Overview of Selection on Observables

Identification Assumptions

Subclassification and Matching

Propensity Scores

Regression

Identification of ATE Under Selection on Observables

Identification Assumption

1. $(Y_1, Y_0) \perp\!\!\!\perp D|X$ (selection on observables)
2. $0 < \mathbb{P}\left[(\right] D = 1|X) < 1$ with probability one (common support)

Identification of ATE Under Selection on Observables

Identification Assumption

1. $(Y_1, Y_0) \perp\!\!\!\perp D|X$ (selection on observables)
2. $0 < \mathbb{P}\left[(\right] D = 1|X) < 1$ with probability one (common support)

Identification Result

Given selection on observables we have

$$\begin{aligned} \mathbb{E}[Y_1 - Y_0|X] &= \mathbb{E}[Y_1 - Y_0|X, D = 1] \\ &= \mathbb{E}[Y|X, D = 1] - \mathbb{E}[Y|X, D = 0] \end{aligned}$$

Therefore, under the common support condition:

$$\begin{aligned} \tau_{ATE} &= \mathbb{E}[Y_1 - Y_0] = \int \mathbb{E}[Y_1 - Y_0|X]\, dP(X) \\ &= \int \left(\mathbb{E}[Y|X, D = 1] - \mathbb{E}[Y|X, D = 0]\right) dP(X) \end{aligned}$$

Discussing Assumptions

- What do these assumptions imply/require?
- Knowledge about the *exact* assignment mechanism, which can be a collection of variables.
- Ability to collect data on these *stratifying* variables.
- Having control and treatment observations in each stratum of the data.
- Most social science data: finite amount of data, self-selection into programs
- This is hard, but not impossible!

Identification of ATT Under Selection on Observables

Identification Assumption

1. $Y_0 \perp\!\!\!\perp D|X$ (selection on observables for controls)
2. $\mathbb{P}\left[(\right] D = 1|X) < 1$ (weak overlap)

Identification Result

Similarly,

$$\begin{aligned} \tau_{ATT} &= \mathbb{E}[Y_1 - Y_0|D = 1] \\ &= \int \big(\mathbb{E}[Y|X, D = 1] - \mathbb{E}[Y|X, D = 0]\big)\, dP(X|D = 1) \end{aligned}$$

Roadmap

Overview of Selection on Observables

Identification Assumptions

Subclassification and Matching

Propensity Scores

Regression

ATT with a Single Covariate

Unit	Potential Outcome Under Treatment	Potential Outcome Under Control	Treatment Status	Covariate
i	Y_{1i}	Y_{0i}	D_i	X_i
1	2	?	1	1
2	4	?	1	1
3	5	?	1	2
4	7	?	1	2
5	6	?	1	2
6		3	0	1
7		5	0	1
8		2	0	2
9		1	0	1
10		0	0	2

ATT with a Single Covariate

Unit	Potential Outcome Under Treatment	Potential Outcome Under Control	Treatment Status	Covariate
i	Y_{1i}	Y_{0i}	D_i	X_i
1	2	?	1	1
2	4	?	1	1
3	5	?	1	2
4	7	?	1	2
5	6	?	1	2
6		3	0	1
7		5	0	1
8		2	0	2
9		1	0	1
10		0	0	2

Subclassification: $\hat{\tau}_{att} = p_1\{(2+4)/2 - (3+5+1)/3\} + p_2\{(5+7+6)/3 - (2+0)/2\}$

ATT with a Single Covariate

Unit	Potential Outcome Under Treatment	Potential Outcome Under Control	Treatment Status	Covariate
i	Y_{1i}	Y_{0i}	D_i	X_i
1	2	?	1	1
2	4	?	1	1
3	5	?	1	2
4	7	?	1	2
5	6	?	1	2
6		3	0	1
7		5	0	1
8		2	0	2
9		1	0	1
10		0	0	2

Subclassification: $\hat{\tau}_{att} = p_1\{(2+4)/2 - (3+5+1)/3\} + p_2\{(5+7+6)/3 - (2+0)/2\}$ where $p_1 = 2/5$ and $p_2 = 3/5$, so $\Longrightarrow \hat{\tau}_{att} = 3$.

ATT with a Single Covariate

Unit	Potential Outcome Under Treatment	Potential Outcome Under Control	Treatment Status	Covariate
i	Y_{1i}	Y_{0i}	D_i	X_i
1	2	?	1	1
2	4	?	1	1
3	5	?	1	2
4	7	?	1	2
5	6	?	1	2
6		3	0	1
7		5	0	1
8		2	0	2
9		1	0	1
10		0	0	2

Subclassification: $\hat{\tau}_{att} = p_1\{(2+4)/2 - (3+5+1)/3\} + p_2\{(5+7+6)/3 - (2+0)/2\}$ where $p_1 = 2/5$ and $p_2 = 3/5$, so $\implies \hat{\tau}_{att} = 3$. What would change in our calculation if we wanted the ATE?

Subclassification

- In subclassification, we partition the data by its covariate values X.
  - E.g,. age, location, race, sex, etc.
- Then, calculate the difference between treated and control within each subclass.
- Finally, take a weighted average of those subclass-specific ATE's, weighted by the distribution of X.

Subclassification, formally

Given that there are K different classes (cells of X), the subclassification estimator for ATE and ATT is

$$\hat{\tau}_{\text{ate}} = \sum_{j=1}^{J} (\bar{Y}_1^j - \bar{Y}_0^j) \frac{N^j}{N}$$

$$\hat{\tau}_{\text{att}} = \sum_{j=1}^{J} (\bar{Y}_1^j - \bar{Y}_0^j) \frac{N_1^j}{N_1}$$

Subclassification, formally

Given that there are K different classes (cells of X), the subclassification estimator for ATE and ATT is

$$\hat{\tau}_{\text{ate}} = \sum_{j=1}^{J} (\bar{Y}_1^j - \bar{Y}_0^j) \frac{N^j}{N}$$

$$\hat{\tau}_{\text{att}} = \sum_{j=1}^{J} (\bar{Y}_1^j - \bar{Y}_0^j) \frac{N_1^j}{N_1}$$

Take the average Y_1 within each class, subtract from it the average Y_0 within the class, then average over classes.

Subclassification, formally

Given that there are K different classes (cells of X), the subclassification estimator for ATE and ATT is

$$\hat{\tau}_{\text{ate}} = \sum_{j=1}^{J} (\bar{Y}_1^j - \bar{Y}_0^j) \frac{N^j}{N}$$

$$\hat{\tau}_{\text{att}} = \sum_{j=1}^{J} (\bar{Y}_1^j - \bar{Y}_0^j) \frac{N_1^j}{N_1}$$

Take the average Y_1 within each class, subtract from it the average Y_0 within the class, then average over classes.

What is the imputed counterfactual control value for treated unit i in group k using this estimator?

Subclassification, formally

Given that there are K different classes (cells of X), the subclassification estimator for ATE and ATT is

$$\hat{\tau}_{\text{ate}} = \sum_{j=1}^{J} (\bar{Y}_1^j - \bar{Y}_0^j) \frac{N^j}{N}$$

$$\hat{\tau}_{\text{att}} = \sum_{j=1}^{J} (\bar{Y}_1^j - \bar{Y}_0^j) \frac{N_1^j}{N_1}$$

Take the average Y_1 within each class, subtract from it the average Y_0 within the class, then average over classes.

What is the imputed counterfactual control value for treated unit i in group k using this estimator? It's the observed mean of the control units in group k.

Matching

- Subclassification breaks down when you have many X's because of the curse of dimensionality – end up with very few units in each cell.
- Exact matching runs into that problem too, but by defining a *distance* between every pair of observations, you know how far off your matches are.

Matching

- Subclassification breaks down when you have many X's because of the curse of dimensionality – end up with very few units in each cell.
- Exact matching runs into that problem too, but by defining a *distance* between every pair of observations, you know how far off your matches are.
- For each treatment unit, we find (a) control unit(s) similar to it (where 'similar' is defined by the distance metric), and trim the sample of any non-matched control units.
- Using the matched dataset, use some method to estimate the treatment effect of interest.
- Goal is different from OLS: we are trying to zoom in on the causal relationship between D and Y by balancing X across D nonparametrically.

Matching

In practice, there are a lot of choices to make when using matching:

- What covariates to include?

Matching

In practice, there are a lot of choices to make when using matching:

- What covariates to include?
- What distance metric to use?

Matching

In practice, there are a lot of choices to make when using matching:

- What covariates to include?
- What distance metric to use?
- Match with or without replacement?

Matching

In practice, there are a lot of choices to make when using matching:

- What covariates to include?
- What distance metric to use?
- Match with or without replacement?
- Match each treatment unit to one or multiple control units?

Matching

In practice, there are a lot of choices to make when using matching:

- What covariates to include?
- What distance metric to use?
- Match with or without replacement?
- Match each treatment unit to one or multiple control units?
- What method do you use to estimate the treatment effect after matching?

Choosing the right covariates is the most important.

Matching

Denote $Y_{j(i)}$ as the imputed counterfactual value for unit i, determined by matching. The standard difference-in-means matching estimator is

$$\hat{\tau}_{att} = \frac{1}{N_1} \sum_{i:D_i=1} \left[(Y_i - \sum_{j=1} \omega_{i,j} Y_{j(i)}) \right]$$

- $\omega_{i,j}$ is determined by the matching procedure that you use: difference in propensity score, Euclidean distance, Mahalanobis distance measure

Issues with Matching: Bias Correction

Especially when matching on many X's, the standard matching estimator can exhibit bias of order $O(N^{-1/k})$ where k is the number of continuous covariates. We can adjust for covariates by estimating a regression in the control group, then predicting based on the covariates what the treated group potential outcome under control would be.

Issues with Matching: Bias Correction

Especially when matching on many X's, the standard matching estimator can exhibit bias of order $O(N^{-1/k})$ where k is the number of continuous covariates. We can adjust for covariates by estimating a regression in the control group, then predicting based on the covariates what the treated group potential outcome under control would be.

The Abadie-Imbens bias adjustment is

$$\tilde{\tau}_{att} = \frac{1}{N_1} \sum_{i:D_i=1} (Y_i - Y_{j(i)}) - (\hat{\mu}_0(X_i) - \hat{\mu}_0(X_{j(i)})),$$

where $\mu(x) = \mathbb{E}[Y \mid X = x, D = 0]$ is the population regression function relating Y_0 to X and $\hat{\mu}(\cdot)$ is an estimate of $\mu(\cdot)$.

Issues with Matching: Bias Correction

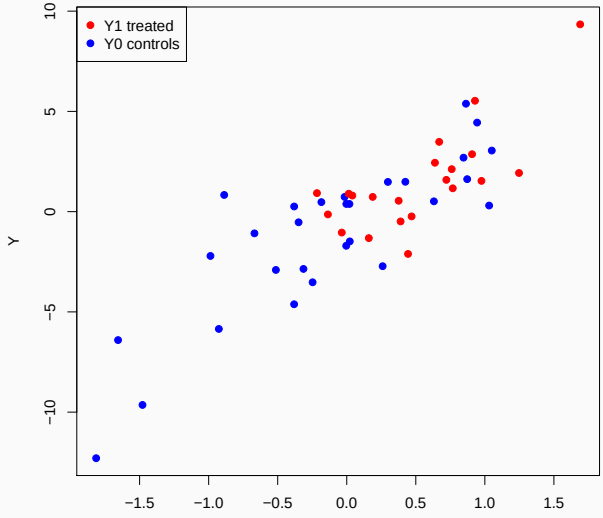
Especially when matching on many X's, the standard matching estimator can exhibit bias of order $O(N^{-1/k})$ where k is the number of continuous covariates. We can adjust for covariates by estimating a regression in the control group, then predicting based on the covariates what the treated group potential outcome under control would be.

The Abadie-Imbens bias adjustment is

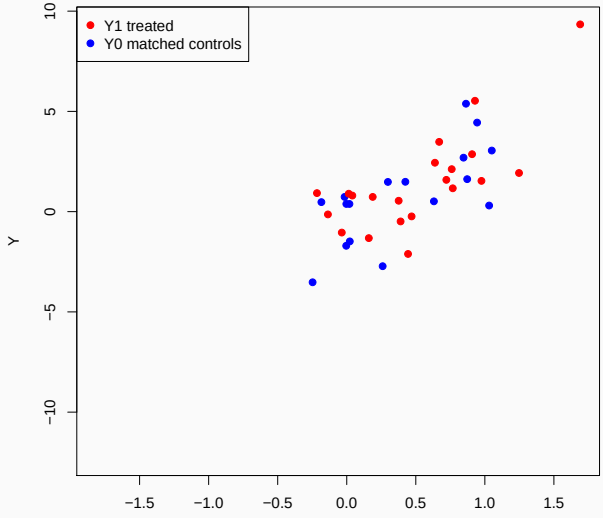
$$\tilde{\tau}_{att} = \frac{1}{N_1} \sum_{i:D_i=1} (Y_i - Y_{j(i)}) - (\hat{\mu}_0(X_i) - \hat{\mu}_0(X_{j(i)})),$$

where $\mu(x) = \mathbb{E}[Y \mid X = x, D = 0]$ is the population regression function relating Y_0 to X and $\hat{\mu}(\cdot)$ is an estimate of $\mu(\cdot)$.

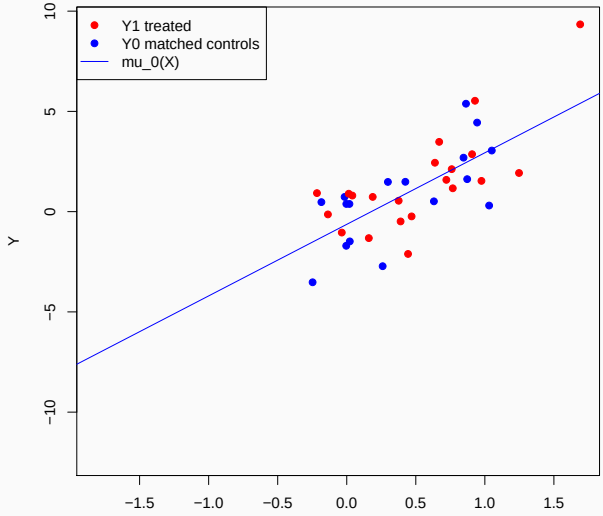
Before Matching



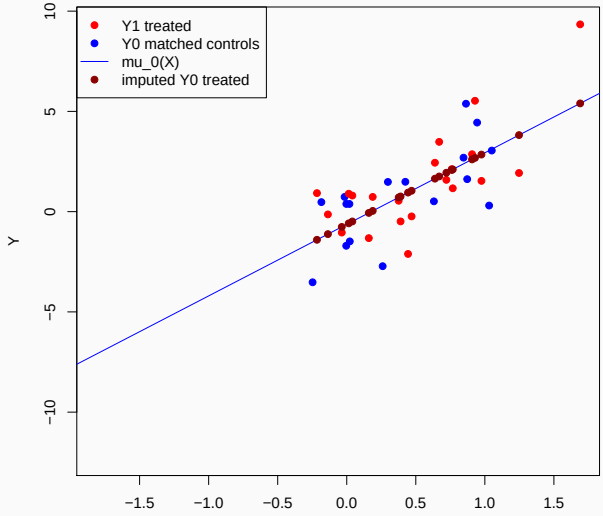
After Matching



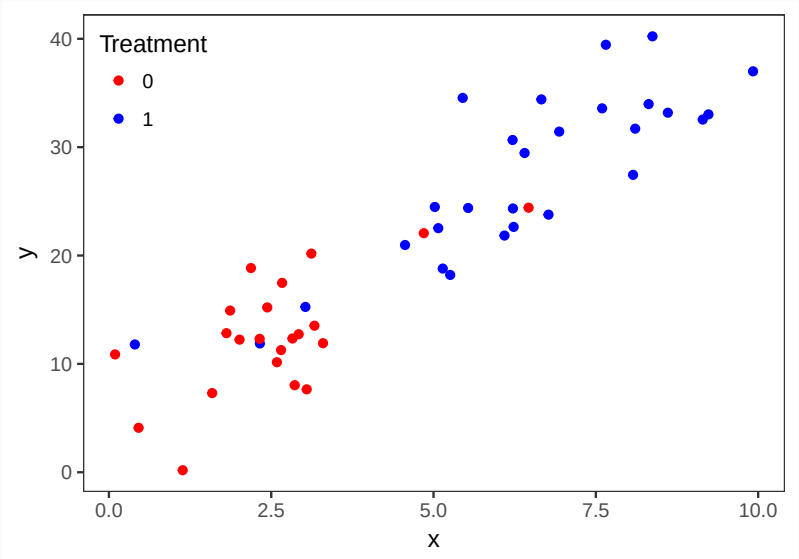
After Matching: Imputatation Function



After Matching: Imputed Unobserved Y_0



Issues with Matching: Not Much Overlap



Matching Considerations

Advantages:

- Good tool for creating balanced samples of treatment and control groups to isolate the treatment effect.
- Easy to use and intuitive interpretation of results

Disadvantages:

- You can get as many results as the number of matching estimators. Choose the one that gets you the best balance.
- Bias-variance trade-off.

Roadmap

Overview of Selection on Observables

Identification Assumptions

Subclassification and Matching

Propensity Scores

Regression

Propensity Scores

The propensity score is simply the probability of treatment, conditional on the covariates, or $\mathbb{P}(D_i = 1|X_i)$. For example:

- In a randomized experiment where half the units are treated, $\mathbb{P}(D_i = 1|X_i) = .5$
- If we treat 30% of group 1 and 80% of group 2, then $\mathbb{P}(D_i|X_i = 1) = .3$ and $\mathbb{P}(D_i|X_i = 2) = .8$.

Propensity Scores

The propensity score is simply the probability of treatment, conditional on the covariates, or $\mathbb{P}(D_i = 1|X_i)$. For example:

- In a randomized experiment where half the units are treated, $\mathbb{P}(D_i = 1|X_i) = .5$
- If we treat 30% of group 1 and 80% of group 2, then $\mathbb{P}(D_i|X_i = 1) = .3$ and $\mathbb{P}(D_i|X_i = 2) = .8$.

If we know the propensity score, we can compare units that have a comparable probability of getting treated and get unbiased treatment effect estimates.

Propensity Score Identification Result

The propensity score is defined as the probability a unit is treated, conditional on confounding covariates:

$$\pi_i := \mathbb{P}(D_i = 1|X)$$

Propensity Score Identification Result

The propensity score is defined as the probability a unit is treated, conditional on confounding covariates:

$$\pi_i := \mathbb{P}(D_i = 1|X)$$

Under the essumtion of selection on observables and common support, the potential outcomes are independent of the treatment assignment. Formally, if we assume

- $(Y_{1i}, Y_{0i}) \perp\!\!\!\perp D_i|X_i$ (selection on observables)
- $0 < \mathbb{P}(D_i = 1|X) < 1$ with probability one (common support)

then we have $(Y_{1i}, Y_{0i}) \perp\!\!\!\perp D \mid \pi_i$. This is called the **Balancing Property** of the propensity score [Rosenbaum and Rubin (1983)]

Propensity Score: So What?

Why do we care about propensity scores?

- Propensity score encodes all the information about the joint dependence of X and D.
- The curse of dimensionality means that as the dimension of X increases, all other points tend to be equally far away.
- The identification result implies that we don't need to condition on *all* the covariates; we can just condition on the propensity score.

Now What?

After estimating the propensity scores, you can:

- Use subclassification or matching based on the estimated propensity score. Important to check balance of the observed covariates; since we don't know the true propensity score, we can check out model by looking at balance.

Now What?

After estimating the propensity scores, you can:

- Use subclassification or matching based on the estimated propensity score. Important to check balance of the observed covariates; since we don't know the true propensity score, we can check out model by looking at balance.
- Weight by the inverse of the propensity score – for example

$$\hat{\tau}_{\text{ate}} = \frac{1}{N}\Big[\sum_{i=1}^{N} \frac{D_i Y_i}{\hat{\pi}(X_i)} - \frac{(1-D_i)Y_i}{1-\hat{\pi}(X_i)}\Big]$$

$$\hat{\tau}_{\text{att}} = \sum_{i=1}^{n} \frac{Y_i(D-\widehat{\pi}(X_i))}{D_i(1-\widehat{\pi}(X_i))}$$

Idea: correct for non-random sampling of units into treatment and control groups by weighting by the reciprocal of the probability of selection into each group.

What could go wrong?

Propensity score weighting assumes that we have uncovered the true propensity score

- In reality, we rarely ever know the true propensity score.
- You may not have all the variables that determine selection into treatment.
- Even if you do have all the variables, it is possible that you have misspecified the model for the propensity score.
- The goal is to create balanced groups of treatment and control, which means you might have to iterate over different models.
- Alternatively, incorporate balance into estimating weights directly [Hainmueller (2012), Imai and Ratkovic (2013)]

Roadmap

Overview of Selection on Observables

Identification Assumptions

Subclassification and Matching

Propensity Scores

Regression

Regression as a type of propensity weighting (Angrist and Pischke, p.83)

Let's our estimand for the ATE comes from a regression of the form

$$\mathbb{E}[Y_i|D_i, X_i] = \alpha + X_i^T \beta + \delta_R D_i$$

where X_i is a vector of covariates and D_i is the treatment indicator.

Regression as a type of propensity weighting (Angrist and Pischke, p.83)

Let's our estimand for the ATE comes from a regression of the form

$$\mathbb{E}[Y_i|D_i, X_i] = \alpha + X_i^T\beta + \delta_R D_i$$

where X_i is a vector of covariates and D_i is the treatment indicator. Using our partialling out formula, we can write δ_R as

$$\delta_R = \frac{\text{Cov}(\tilde{D}_i, Y_i)}{\text{Var}(\tilde{D}_i)}$$

Regression as a type of propensity weighting (Angrist and Pischke, p.83)

Let's our estimand for the ATE comes from a regression of the form

$$\mathbb{E}[Y_i|D_i, X_i] = \alpha + X_i^T\beta + \delta_R D_i$$

where X_i is a vector of covariates and D_i is the treatment indicator. Using our partialling out formula, we can write δ_R as

$$\delta_R = \frac{\text{Cov}(\tilde{D}_i, Y_i)}{\text{Var}(\tilde{D}_i)} = \frac{\mathbb{E}[(D_i - \mathbb{E}[D_i|X_i])Y_i]}{\mathbb{E}[(D_i - \mathbb{E}[D_i|X_i])^2]}$$

Regression as a type of propensity weighting (Angrist and Pischke, p.83)

Let's our estimand for the ATE comes from a regression of the form

$$\mathbb{E}[Y_i|D_i, X_i] = \alpha + X_i^T\beta + \delta_R D_i$$

where X_i is a vector of covariates and D_i is the treatment indicator. Using our partialling out formula, we can write δ_R as

$$\delta_R = \frac{\mathsf{Cov}(\tilde{D}_i, Y_i)}{\mathsf{Var}(\tilde{D}_i)} = \frac{\mathbb{E}[(D_i - \mathbb{E}[D_i|X_i])Y_i]}{\mathbb{E}[(D_i - \mathbb{E}[D_i|X_i])^2]} = \frac{\mathbb{E}[(D_i - \mathbb{E}[D_i|X_i])Y_i]}{\mathbb{E}[\mathbb{E}[(D_i - \mathbb{E}[D_i|X_i])^2|X_i]]}$$

Regression as a type of propensity weighting (Angrist and Pischke, p.83)

Let's our estimand for the ATE comes from a regression of the form

$$\mathbb{E}[Y_i|D_i, X_i] = \alpha + X_i^T\beta + \delta_R D_i$$

where X_i is a vector of covariates and D_i is the treatment indicator. Using our partialling out formula, we can write δ_R as

$$\begin{aligned}\delta_R &= \frac{\mathsf{Cov}(\tilde{D}_i, Y_i)}{\mathsf{Var}(\tilde{D}_i)} = \frac{\mathbb{E}[(D_i - \mathbb{E}[D_i|X_i])Y_i]}{\mathbb{E}[(D_i - \mathbb{E}[D_i|X_i])^2]} = \frac{\mathbb{E}[(D_i - \mathbb{E}[D_i|X_i])Y_i]}{\mathbb{E}[\mathbb{E}[(D_i - \mathbb{E}[D_i|X_i])^2|X_i]]} \\ &= \frac{\mathbb{E}[(D_i - \mathbb{E}[D_i|X])Y_i]}{\mathbb{E}[\mathsf{Var}[D_i|X]]}\end{aligned}$$

Regression as a type of propensity weighting (Angrist and Pischke, p.83)

Let's our estimand for the ATE comes from a regression of the form

$$\mathbb{E}[Y_i|D_i, X_i] = \alpha + X_i^T\beta + \delta_R D_i$$

where X_i is a vector of covariates and D_i is the treatment indicator. Using our partialling out formula, we can write δ_R as

$$\begin{aligned}\delta_R &= \frac{\mathsf{Cov}(\tilde{D}_i, Y_i)}{\mathsf{Var}(\tilde{D}_i)} = \frac{\mathbb{E}[(D_i - \mathbb{E}[D_i|X_i])Y_i]}{\mathbb{E}[(D_i - \mathbb{E}[D_i|X_i])^2]} = \frac{\mathbb{E}[(D_i - \mathbb{E}[D_i|X_i])Y_i]}{\mathbb{E}[\mathbb{E}[(D_i - \mathbb{E}[D_i|X_i])^2|X_i]]} \\ &= \frac{\mathbb{E}[(D_i - \mathbb{E}[D_i|X])Y_i]}{\mathbb{E}[\mathsf{Var}[D_i|X]]} = \frac{\mathbb{E}[(D_i - \mathbb{E}[D_i|X])Y_i]}{\mathbb{E}[\mathbb{E}[D_i|X_i](1 - \mathbb{E}[D_i|X_i])]}\end{aligned}$$

Regression as a type of propensity weighting

Now let's substitute in the definition of the propensity score function, $p(X_i) = \mathbb{E}[D_i|X_i]$.

$$\delta_R = \frac{\mathbb{E}[(D_i - \mathbb{E}[D_i|X])Y_i]}{\mathbb{E}[\mathbb{E}[D_i|X_i](1 - \mathbb{E}[D_i|X_i])]} = \frac{\mathbb{E}[(D_i - p(X_i))Y_i]}{\mathbb{E}[p(X_i)(1 - p(X_i))]}$$

Regression as a type of propensity weighting

Now let's substitute in the definition of the propensity score function, $p(X_i) = \mathbb{E}[D_i|X_i]$.

$$\delta_R = \frac{\mathbb{E}[(D_i - \mathbb{E}[D_i|X])Y_i]}{\mathbb{E}[\mathbb{E}[D_i|X_i](1 - \mathbb{E}[D_i|X_i])]} = \frac{\mathbb{E}[(D_i - p(X_i))Y_i]}{\mathbb{E}[p(X_i)(1 - p(X_i))]}$$

We can see that this estimand is equal to the weighted propensity score estimand:

$$\delta_R = \mathbb{E}\Big[\frac{p(X_i)(1 - p(X_i))}{\mathbb{E}[p(X_i)(1 - p(X_i))]}\Big(\frac{Y_i D_i}{p(X_i)} - \frac{Y_i(1 - D_i)}{1 - p(X_i)}\Big)\Big]$$

where $\frac{p(X_i)(1-p(X_i))}{\mathbb{E}[p(X_i)(1-p(X_i))]}$ is the weight for observations with covariates X_i.

Regression as a type of propensity weighting

Now let's substitute in the definition of the propensity score function, $p(X_i) = \mathbb{E}[D_i|X_i]$.

$$\delta_R = \frac{\mathbb{E}[(D_i - \mathbb{E}[D_i|X])Y_i]}{\mathbb{E}[\mathbb{E}[D_i|X_i](1 - \mathbb{E}[D_i|X_i])]} = \frac{\mathbb{E}[(D_i - p(X_i))Y_i]}{\mathbb{E}[p(X_i)(1 - p(X_i))]}$$

We can see that this estimand is equal to the weighted propensity score estimand:

$$\delta_R = \mathbb{E}\Big[\frac{p(X_i)(1-p(X_i))}{\mathbb{E}[p(X_i)(1-p(X_i))]}\Big(\frac{Y_i D_i}{p(X_i)} - \frac{Y_i(1-D_i)}{1-p(X_i)}\Big)\Big]$$

where $\frac{p(X_i)(1-p(X_i))}{\mathbb{E}[p(X_i)(1-p(X_i))]}$ is the weight for observations with covariates X_i. Compare this to the unweighted propensity score estimand:

$$\delta_{ATE} = \mathbb{E}\Big[\frac{Y_i D_i}{p(X_i)} - \frac{Y_i(1-D_i)}{1-p(X_i)}\Big]$$

When will these two coincide?

Regression as a type of propensity weighting

Now let's substitute in the definition of the propensity score function, $p(X_i) = \mathbb{E}[D_i|X_i]$.

$$\delta_R = \frac{\mathbb{E}[(D_i - \mathbb{E}[D_i|X])Y_i]}{\mathbb{E}[\mathbb{E}[D_i|X_i](1 - \mathbb{E}[D_i|X_i])]} = \frac{\mathbb{E}[(D_i - p(X_i))Y_i]}{\mathbb{E}[p(X_i)(1 - p(X_i))]}$$

We can see that this estimand is equal to the weighted propensity score estimand:

$$\delta_R = \mathbb{E}\Big[\frac{p(X_i)(1 - p(X_i))}{\mathbb{E}[p(X_i)(1 - p(X_i))]}\Big(\frac{Y_i D_i}{p(X_i)} - \frac{Y_i(1 - D_i)}{1 - p(X_i)}\Big)\Big]$$

where $\frac{p(X_i)(1-p(X_i))}{\mathbb{E}[p(X_i)(1-p(X_i))]}$ is the weight for observations with covariates X_i. Compare this to the unweighted propensity score estimand:

$$\delta_{ATE} = \mathbb{E}\Big[\frac{Y_i D_i}{p(X_i)} - \frac{Y_i(1 - D_i)}{1 - p(X_i)}\Big]$$

When will these two coincide? Constant treatment effects across strata of X_i. Otherwise, OLS does not estimate ATE/ATT.

OLS with heterogeneous treatment effects

In this case your estimat**or** is picking your estim**and**.

Hybrid Approaches: AIPW Estimator

$$\widehat{\tau}_{DR} = \frac{1}{N}\sum_{i=1}^{n}\left(\frac{D_i(Y_i - \hat{\mu}_1(X_i))}{\hat{\pi}(X_i)} - \frac{(1-D_i)(Y_i - \hat{\mu}_0(X_i))}{1-\hat{\pi}(X_i)} + \{\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)\}\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\underbrace{\left[\overbrace{\hat{\mu}_1(\boldsymbol{X}_i)}^{\text{Regression}} + \overbrace{\frac{D_i(Y_i - \hat{\mu}_1(\boldsymbol{X}_i))}{\hat{\pi}(\boldsymbol{X}_i)}}^{\text{IPW}}\right]}_{\text{estimator for } \mathbb{E}[Y_i(1)]} - \frac{1}{n}\sum_{i=1}^{n}\underbrace{\left[\hat{\mu}_0(\boldsymbol{X}_i) + \frac{(1-D_i)(Y_i - \hat{\mu}_0(\boldsymbol{X}_i))}{1-\hat{\pi}(\boldsymbol{X}_i)}\right]}_{\text{estimator for } \mathbb{E}[Y_i(0)]}$$

Fit $\widehat{\mu}, \widehat{\pi}$ using learner of choice. More to come in 450C.