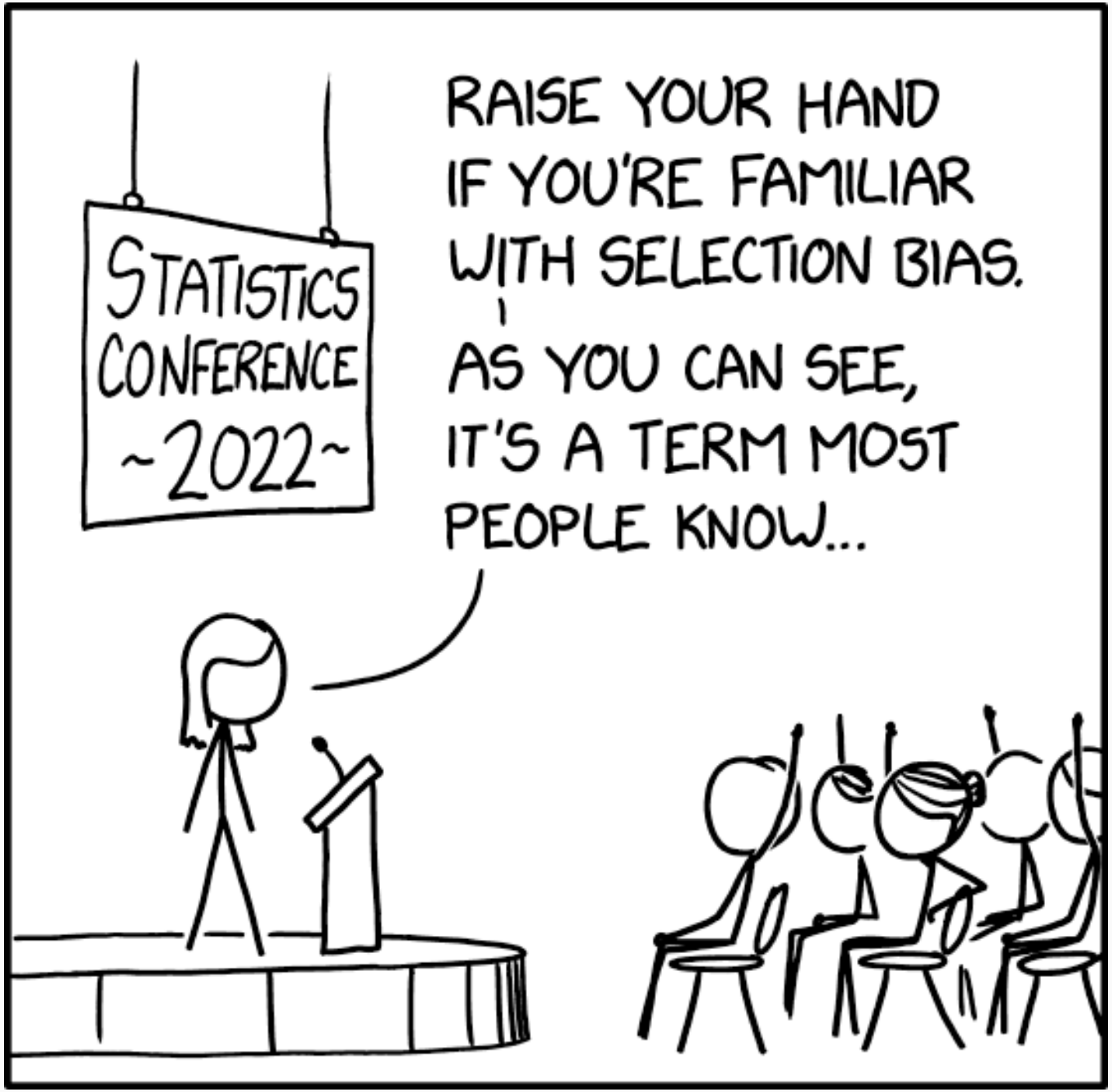
Impact Evaluation Methods

Topic 4: Matching

Alex Alekseev

University of Regensburg, Department of Economics


STATISTICS CONFERENCE ~2022~
RAISE YOUR HAND IF YOU'RE FAMILIAR WITH SELECTION BIAS.
AS YOU CAN SEE, IT'S A TERM MOST PEOPLE KNOW...

Previously on *Impact Evaluation Methods*...

- Causal diagrams (aka directed acyclical graphs)
- Three strategies
  - Conditioning
  - Instrumental variables
  - Front-door method
- Back-door criterion
- How to find back-door paths

Stratification

Set-Up

- Back to the language of the potential outcomes framework
- The causal effect of a binary variable D on an outcome variable Y
- We do not wish to assume independence $((Y^1, Y^0) \perp D)$
- However, we are willing to assume **conditional independence**, where conditioning is with respect to a set of observable variables S:

$$(Y^1, Y^0) \perp D \mid S$$

- Conditional on S, the potential outcomes do not depend on treatment exposure

Common support

- We are also going to assume **common support**

$$0 < \mathbb{P}(D = 1 \mid S) < 1.$$

- This assumption rules out cases when for some values of S no one can be in the treatment state or no one is in the control state

Implications

- These assumptions imply that

$$\mathbb{E}[Y^0 \mid D = 1, S] = \mathbb{E}[Y^0 \mid D = 0, S]$$
$$\mathbb{E}[Y^1 \mid D = 1, S] = \mathbb{E}[Y^1 \mid D = 0, S],$$

- The potential outcomes in the treatment and control states do not depend on D once we condition on S

Note

However, that these assumption **do not** imply that, e.g.,

$$\mathbb{E}[Y^0 \mid D = 1, S] = \mathbb{E}[Y^1 \mid D = 1, S]$$

Stratification

- Assume that variables in S are discrete (or can be discretized)
- Knowledge and observation of S allow for a "perfect stratification" of the data
- Individuals within groups defined by values of the variables in S are indistinguishable from each other in all ways except
  - observed treatment status
  - differences in the potential outcomes that are independent of treatment status
- The naive treatment effect will be biased relative to the average treatment effect
- The bias will disappear after conditioning on S
- **Treatment assignment is ignorable** or **treatment selection is on observables**

Conditioning

- Conditioning on S allows us to derive the following identities:

$$
\begin{aligned}
NTE(S) &= \mathbb{E}[Y^1 \mid D = 1, S] - \mathbb{E}[Y^0 \mid D = 0, S] \\
ATT(S) &= \mathbb{E}[\delta \mid D = 1, S] \\
&= \mathbb{E}[Y^1 \mid D = 1, S] - \mathbb{E}[Y^0 \mid D = 1, S] \\
&= \mathbb{E}[Y^1 \mid D = 1, S] - \mathbb{E}[Y^0 \mid D = 0, S] \\
ATU(S) &= \mathbb{E}[\delta \mid D = 0, S] \\
&= \mathbb{E}[Y^1 \mid D = 0, S] - \mathbb{E}[Y^0 \mid D = 0, S] \\
&= \mathbb{E}[Y^1 \mid D = 1, S] - \mathbb{E}[Y^0 \mid D = 0, S] \\
ATE(S) &= \mathbb{E}[\delta \mid S] \\
&= \mathbb{P}(D = 1)ATT(S) + (1 - \mathbb{P}(D = 1))ATU(S) \\
&= \mathbb{E}[Y^1 \mid D = 1, S] - \mathbb{E}[Y^0 \mid D = 0, S].
\end{aligned}
$$

- Conditional on S, all four treatment effects are identical

Note

This does not mean, however, that the unconditional $NTE = \mathbb{E}[Y^1 \mid D = 1] - \mathbb{E}[Y^1 \mid D = 0]$ is unbiased.

Unconditional Effects

- How do we get from conditional effects to unconditional ones?
- By appropriately weighting the conditional effects

$$\mathbb{E}[\delta] = \mathbb{E}_S[\mathbb{E}[\delta \mid S]]$$
$$\mathbb{E}[\delta \mid D = 1] = \mathbb{E}_{S|D=1}[\mathbb{E}[\delta \mid D = 1, S]]$$
$$\mathbb{E}[\delta \mid D = 0] = \mathbb{E}_{S|D=0}[\mathbb{E}[\delta \mid D = 0, S]].$$

Math Time: ATT

- Let's have a closer look at how the conditional independence assumptions allows us to compute those expectations
- For example, for the ATT we have

$$\begin{aligned} ATT &= \mathbb{E}[Y^1 \mid D = 1] - \mathbb{E}[Y^0 \mid D = 1] \\ &= \sum_{s=1}^{\bar{S}} \mathbb{E}[Y^1 \mid D = 1, S = s] \mathbb{P}(S = s \mid D = 1) \\ &\quad - \sum_{s=1}^{\bar{S}} \mathbb{E}[Y^0 \mid D = 1, S = s] \mathbb{P}(S = s \mid D = 1) \end{aligned}$$

Math Time: ATT

The conditional independence assumption allows us to substitute counterfactual terms $\mathbb{E}[Y^0 \mid D = 1, S = s]$ with observed terms $\mathbb{E}[Y^0 \mid D = 0, S = s]$

$$
\begin{aligned}
ATT &= \sum_{s=1}^{\bar{S}} \mathbb{E}[Y^1 \mid D = 1, S = s] \mathbb{P}(S = s \mid D = 1) \\
&\quad - \sum_{s=1}^{\bar{S}} \mathbb{E}[Y^0 \mid D = 0, S = s] \mathbb{P}(S = s \mid D = 1) \\
&= \sum_{s=1}^{\bar{S}} \left(\mathbb{E}[Y^1 \mid D = 1, S = s] - \mathbb{E}[Y^0 \mid D = 0, S = s] \right) \mathbb{P}(S = s \mid D = 1) \\
&= \sum_{s=1}^{\bar{S}} NTE(S = s) \mathbb{P}(S = s \mid D = 1).
\end{aligned}
$$

ATT and Matching

- Notice the term $\sum_{s=1}^{\bar{S}} \mathbb{E}[Y^0 \mid D = 0, S = s]\mathbb{P}(S = s \mid D = 1)$
- We are taking the expectation not with respect to the actual distribution $\mathbb{P}(S = s \mid D = 0)$, but with respect to the distribution $\mathbb{P}(S = s \mid D = 1)$
- This ensures that the conditional distribution of S given $D = 1$ is identical between the treatment and control units
- You can think of it as weighting the control units in a way that achieves the covariate balance between the treatment and control groups
- This is the main purpose of **matching**
- We are weighting the control units to look like the treatment units in terms of S

Math Time: ATU

$$
\begin{aligned}
ATU &= \mathbb{E}[Y^1 \mid D = 0] - \mathbb{E}[Y^0 \mid D = 0] \\
&= \sum_{s=1}^{\bar{S}} \mathbb{E}[Y^1 \mid D = 0, S = s]\mathbb{P}(S = s \mid D = 0) \\
&\quad - \sum_{s=1}^{\bar{S}} \mathbb{E}[Y^0 \mid D = 0, S = s]\mathbb{P}(S = s \mid D = 0) \\
&= \sum_{s=1}^{\bar{S}} \mathbb{E}[Y^1 \mid D = 1, S = s]\mathbb{P}(S = s \mid D = 0) \\
&\quad - \sum_{s=1}^{\bar{S}} \mathbb{E}[Y^0 \mid D = 0, S = s]\mathbb{P}(S = s \mid D = 0) \\
&= \sum_{s=1}^{\bar{S}} \Big(\mathbb{E}[Y^1 \mid D = 1, S = s] - \mathbb{E}[Y^0 \mid D = 0, S = s]\Big)\, \mathbb{P}(S = s \mid D = 0) \\
&= \sum_{s=1}^{\bar{S}} NTE(S = s)\mathbb{P}(S = s \mid D = 0).
\end{aligned}
$$

ATU and Matching

- The ATU is the weighted average of the NTE in each stratum defined by S where the weights are given by the conditional distribution of S given $D = 0$
- Notice the term $\sum_{s=1}^{\bar{S}} \mathbb{E}[Y^1 \mid D = 1, S = s]\mathbb{P}(S = s \mid D = 0)$
- We are taking the expectation not with respect to the actual distribution of $\mathbb{P}(S = s \mid D = 1)$ but with respect to the distribution $\mathbb{P}(S = s \mid D = 0)$
- Ensures the covariate balance between the treatment and control units
- We are weighting the treatment units to look like the control units in terms of S

ATE

- The *ATE* can be simply found as

$$ATE = \sum_{s=1}^{\bar{S}} \mathbb{E}[\delta \mid S]\mathbb{P}(S = s) = \sum_{s=1}^{\bar{S}} NTE(S = s)\mathbb{P}(S = s)$$

Math time: Human capital

- Suppose we are interested in the causal effect of going to college (D) on income (Y)
- The effect is confounded by ability ($S \in \{1, 2, 3\}$) that jointly causes whether someone attends college or not and income

Math time: Human capital

- The distribution of ability in the population and the conditional distribution of D given S are

S	$\mathbb{P}(S)$
1	0.44
2	0.24
3	0.32

	$S=1$	$S=2$	$S=3$
$\mathbb{P}(D=0 \mid S)$	9/11	1/2	3/8
$\mathbb{P}(D=1 \mid S)$	2/11	1/2	5/8

- We can use these two tables to derive the joint distribution $\mathbb{P}(D, S)$:

$$\mathbb{P}(D=d, S=s) = \mathbb{P}(D=d \mid S=s)\mathbb{P}(S=s)$$

	$D=0$	$D=1$	$\mathbb{P}(S)$
$S=1$	0.36	0.08	0.44
$S=2$	0.12	0.12	x
$S=3$	0.12	0.2	x
$\mathbb{P}(D)$	x	x	x

Math time: Human capital

- The distribution of ability in the population and the conditional distribution of D given S are

S	$\mathbb{P}(S)$
1	0.44
2	0.24
3	0.32

	$S=1$	$S=2$	$S=3$
$\mathbb{P}(D=0 \mid S)$	9/11	1/2	3/8
$\mathbb{P}(D=1 \mid S)$	2/11	1/2	5/8

- We can use these two tables to derive the joint distribution $\mathbb{P}(D, S)$:

	$D=0$	$D=1$	$\mathbb{P}(S)$
$S=1$	0.36	0.08	0.44
$S=2$	0.12	0.12	0.24
$S=3$	0.12	0.2	0.32
$\mathbb{P}(D)$	0.6	0.4	1

Math time: Human capital

	$D=0$	$D=1$	$\mathbb{P}(S)$
$S=1$	0.36	0.08	0.44
$S=2$	0.12	0.12	0.24
$S=3$	0.12	0.2	0.32
$\mathbb{P}(D)$	0.6	0.4	1

- Finally, let's derive the conditional distribution of S given D using the Bayes' rule:

$$\mathbb{P}(S=s \mid D=d) = \frac{\mathbb{P}(S=s, D=d)}{\mathbb{P}(D=d)}$$

	$D=0$	$D=1$
$\mathbb{P}(S=1 \mid D)$	0.6	0.2
$\mathbb{P}(S=2 \mid D)$	0.2	0.3
$\mathbb{P}(S=3 \mid D)$	0.2	0.5

Math time: Human capital

	$D = 0$	$D = 1$	$\mathbb{P}(S)$
$S = 1$	0.36	0.08	0.44
$S = 2$	0.12	0.12	0.24
$S = 3$	0.12	0.2	0.32
$\mathbb{P}(D)$	0.6	0.4	1

- Finally, let's derive the conditional distribution of S given D using the Bayes' rule:

$$\mathbb{P}(S = s \mid D = d) = \frac{\mathbb{P}(S = s, D = d)}{\mathbb{P}(D = d)}$$

	$D = 0$	$D = 1$
$\mathbb{P}(S = 1 \mid D)$	0.6	0.2
$\mathbb{P}(S = 2 \mid D)$	0.2	0.3
$\mathbb{P}(S = 3 \mid D)$	0.2	0.5

Math time: Human capital

- The potential outcomes for income are given by the following table
- We are assuming conditional independence

	$\mathbb{E}[Y^0 \mid D=0, S]$	$\mathbb{E}[Y^1 \mid D=1, S]$	$\mathbb{E}[\delta \mid S]$
$S=1$	2	4	2
$S=2$	6	8	2
$S=3$	10	14	4

Math time: Human capital

	$\mathbb{E}[Y^0 \mid D=0, S]$	$\mathbb{E}[Y^1 \mid D=1, S]$	$\mathbb{E}[\delta \mid S]$
$S=1$	2	4	2
$S=2$	6	8	2
$S=3$	10	14	4

	$D=0$	$D=1$
$\mathbb{P}(S=1 \mid D)$	0.6	0.2
$\mathbb{P}(S=2 \mid D)$	0.2	0.3
$\mathbb{P}(S=3 \mid D)$	0.2	0.5

- Now we can easily compute all the treatment effects.

$$ATT = \sum_{s=1}^{\bar{s}} \mathbb{E}[\delta \mid S]\mathbb{P}(S=s \mid D=1) = 3$$

Math time: Human capital

	$\mathbb{E}[Y^0 \mid D=0, S]$	$\mathbb{E}[Y^1 \mid D=1, S]$	$\mathbb{E}[\delta \mid S]$
$S=1$	2	4	2
$S=2$	6	8	2
$S=3$	10	14	4

	$D=0$	$D=1$
$\mathbb{P}(S=1 \mid D)$	0.6	0.2
$\mathbb{P}(S=2 \mid D)$	0.2	0.3
$\mathbb{P}(S=3 \mid D)$	0.2	0.5

- Now we can easily compute all the treatment effects.

$$ATT = \sum_{s=1}^{\bar{S}} \mathbb{E}[\delta \mid S]\mathbb{P}(S=s \mid D=1) = 2\times 0.2 + 2\times 0.3 + 4\times 0.5 = 3$$

Math time: Human capital

	$\mathbb{E}[Y^0 \mid D=0, S]$	$\mathbb{E}[Y^1 \mid D=1, S]$	$\mathbb{E}[\delta \mid S]$
$S=1$	2	4	2
$S=2$	6	8	2
$S=3$	10	14	4

	$D=0$	$D=1$
$\mathbb{P}(S=1 \mid D)$	0.6	0.2
$\mathbb{P}(S=2 \mid D)$	0.2	0.3
$\mathbb{P}(S=3 \mid D)$	0.2	0.5

- Now we can easily compute all the treatment effects.

$$ATU = \sum_{s=1}^{\bar{s}} \mathbb{E}[\delta \mid S]\mathbb{P}(S=s \mid D=0)$$

Math time: Human capital

	$\mathbb{E}[Y^0 \mid D=0, S]$	$\mathbb{E}[Y^1 \mid D=1, S]$	$\mathbb{E}[\delta \mid S]$
$S=1$	2	4	2
$S=2$	6	8	2
$S=3$	10	14	4

	$D=0$	$D=1$
$\mathbb{P}(S=1 \mid D)$	0.6	0.2
$\mathbb{P}(S=2 \mid D)$	0.2	0.3
$\mathbb{P}(S=3 \mid D)$	0.2	0.5

- Now we can easily compute all the treatment effects.

$$ATU = \sum_{s=1}^{\bar{S}} \mathbb{E}[\delta \mid S]\mathbb{P}(S=s \mid D=0) = 2\times 0.6 + 2\times 0.2 + 4\times 0.2 = 2.4$$

Math time: Human capital

	$\mathbb{E}[Y^0 \mid D=0, S]$	$\mathbb{E}[Y^1 \mid D=1, S]$	$\mathbb{E}[\delta \mid S]$
$S=1$	2	4	2
$S=2$	6	8	2
$S=3$	10	14	4

	$D=0$	$D=1$
$\mathbb{P}(S=1 \mid D)$	0.6	0.2
$\mathbb{P}(S=2 \mid D)$	0.2	0.3
$\mathbb{P}(S=3 \mid D)$	0.2	0.5

- Now we can easily compute all the treatment effects.

$$ATE = \sum_{s=1}^{\bar{S}} \mathbb{E}[\delta \mid S]\mathbb{P}(S=s)$$

Math time: Human capital

	$\mathbb{E}[Y^0 \mid D=0, S]$	$\mathbb{E}[Y^1 \mid D=1, S]$	$\mathbb{E}[\delta \mid S]$
$S=1$	2	4	2
$S=2$	6	8	2
$S=3$	10	14	4

	$D=0$	$D=1$
$\mathbb{P}(S=1 \mid D)$	0.6	0.2
$\mathbb{P}(S=2 \mid D)$	0.2	0.3
$\mathbb{P}(S=3 \mid D)$	0.2	0.5

- Now we can easily compute all the treatment effects.

$$ATE = \sum_{s=1}^{\bar{S}} \mathbb{E}[\delta \mid S]\mathbb{P}(S=s) = 2 \times 0.44 + 2 \times 0.24 + 4 \times 0.32 = 2.64$$

Matching

Matching

- The stratification exercise illustrates what matching achieves
- It does not show how matching, as an algorithm, actually works
- Two equivalent ways to implements matching: by **selecting a matched sample** or by **weighting**
- We will first discuss selecting matched samples
- Then show that it can be thought of as a special case of weighting

Exact Matching

- For the ATT, exact matching constructs the counterfactual for each treatment case using the control cases with **identical** values on the variables in S
- If there are several matches, average them or randomly choose one match
- Estimate the ATT as follows

$$\widehat{ATT} = \frac{1}{n^1} \sum_{i \in I} \left(y_i - y_{j(i)} \right)$$

- For the ATU:

$$\widehat{ATU} = \frac{1}{n^0} \sum_{j \in J} \left(y_{i(j)} - y_j \right)$$

ATE

- The ATE can be estimated as

$$\begin{aligned}
\widehat{ATE} &= \frac{n^1}{n^0+n^1}\widehat{ATT} + \frac{n^0}{n^0+n^1}\widehat{ATU} \\
&= \frac{n^1}{n^0+n^1}\frac{1}{n^1}\sum_{i\in I}\left(y_i - y_{j(i)}\right) + \frac{n^0}{n^0+n^1}\frac{1}{n^0}\sum_{j\in J}\left(y_{i(j)} - y_j\right) \\
&= \frac{1}{n^0+n^1}\left[\sum_{i\in I}(2d_i-1)(y_i - y_{j(i)}) + \sum_{j\in J}(2d_j-1)(y_j - y_{i(j)})\right] \\
&= \frac{1}{n}\sum_{k=1}^{n}(2d_k-1)\left(y_k - y_{l(k)}\right),
\end{aligned}$$

Limitations

- If we have a lot of covariates that take many values we will often not be able to find matches
- There may be many strata in the available data in which no treatment or control cases are observed, even though the true probability of being treated is between 0 and 1 for every stratum in the population
- This problem is called **sparseness**

Coarsened Exact Matching

- One way to deal with sparseness while preserving the core idea of exact matching is to use **coarsened exact matching**
- It's based on the idea that sometimes it's possible to do exact matching once we coarsen the data enough
- We coarsen the data by binning the values of continuous (or even categorical) variables, then try to find exact matches

Nearest-Neighbor Matching

- For the ATT, construct the counterfactual for each treatment case using the control cases that are "closest" to the treatment case
- "Closest" on a unidimensional measure ("distance") constructed from the variables in S.
- The algorithm can be run with or without replacement
- With replacement, a control case is returned to the pool after a match and can be matched later to another treatment case
- Without replacement, a control case is taken out of the pool once it is matched

Nearest-Neighbor Matching Types

- One can select only one nearest neighbor for each treatment case (one-to-one matching)...
- ...or match top k nearest neighbors (k-nearest-neighbor matching) to each target treatment case, in which case we average over the matched nearest neighbors
- Matching more control cases to each treatment case results in lower expected variance of the treatment effect estimate
- ...but also raises the possibility of greater bias, because the probability of making more poor matches increases

Poor Matches

- A danger with nearest-neighbor matching is that it may result in some very poor matches for treatment cases
- **Caliper matching** is designed to remedy this drawback by restricting matches to some maximum distance
- With this type of matching, some treatment cases may not receive matches
- **Radius matching** matches all control cases within a particular distance–the "radius"–from the treatment case and gives the selected control cases equal weight
- If there are no control cases within the radius of a particular treatment case, the nearest available control case is used

Mahalanobis Distance

- How can we define the "distance" between cases?
- There are two popular approaches
- The first approach uses the values of covariates to compute the **Mahalanobis distance**, which is a scale-invariant distance metric:

$$d(S_i, S_j) = \sqrt{(S_i - S_j)' \Sigma^{-1} (S_i - S_j)},$$

Propensity Score

- Another popular approach is to use the **propensity score**
- Propensity score matching (Rosenbaum and Rubin, 1983) uses variables in S to estimate the probability of treatment conditional on S, $\mathbb{P}(D = 1 \mid S)$ (e.g., using logit)
- Then uses the predicted values from that estimation to collapse the covariates into a single scalar called the **propensity score**
- All comparisons between the treatment and control group are then based on that value

Note

Although propensity score matching is a popular technique, there is some pushback against using it (King and Nielsen 2019). The general advice is to use propensity scores to construct weights, not select matches.

Weighting

Generalization of Matching

- Exact matching or nearest-neighbor-matching can be thought of as a special case of **weighting**
- For example, the treatment effect for the treated can be defined as

$$\widehat{ATT} = \frac{1}{n^1} \sum_{i \in I} \left[y_i - \sum_{j \in J} w_{ij} y_j \right]$$

- In the case of exact matching the control unit j that is matched to the treatment unit i gets a weight of 1, all others receive a weight of 0
- In case there are several matched control cases, each of them receives an equal weight

ATU and ATE

- The treatment effect for the untreated can be defined as

$$\widehat{ATU} = \frac{1}{n^0} \sum_{j \in J} \left[\sum_{i \in I} w_{ji} y_i - y_j \right]$$

- The average treatment effect can be defined in one of the two ways:

$$\begin{aligned} \widehat{ATE} &= \frac{n^1}{n} \widehat{ATT} + \frac{n^0}{n} \widehat{ATU} \\ &= \frac{1}{n} \left(\sum_{i \in I} w_i y_i - \sum_{j \in J} w_j y_j \right) \end{aligned}$$

Kernel Matching

- In kernel matching, a **kernel function** is used to construct the weights (Heckman, Ichimura, Smith, and Todd (1998), Smith and Todd (2005))
- A kernel function G takes the distance between the two units and, together with a specified **bandwidth**, returns a value
- The value will be the highest if the distance is zero
- All control units are matched to each treatment unit but weighted so that those closest to the treatment unit are given the greatest weight
- The weights of all the control units that are matched to the treatment unit i are given by

$$w_{ij} = \frac{G(d(S_i, S_j))}{\sum_{j \in J} G(d(S_i, S_j))}$$

Inverse Probability Weighting

- Inverse probability weighting uses propensity scores to construct weights (See, e.g., Horvitz and Thompson (1952), Hirano, Imbens, and Ridder (2003) or Busso, DiNardo, and McCrary (2014))
- After estimating the propensity score $p(S) \equiv \mathbb{P}(D = 1 \mid S)$, the weights for the control units in the ATT formula are given by

$$w_{ij} = \frac{1}{n^1} \frac{p_j}{1 - p_j},$$

- For control observations with high propensity scores (high p_j) the weights are going to be high, and vice versa
- Control observations with high propensity scores are basically the treatment observations that did not get treated (counterfactuals)

ATU

- For the ATU, we have the following weights:

$$w_{ji} = \frac{1}{n^0} \frac{1 - p_i}{p_i},$$

- For treatment observations with low propensity scores (low p_i) the weights are going to be high, and vice versa
- Treatment observations with low propensity scores are like the control observations that did get treated (counterfactuals)

Note

The inverse probability weight also has some modifications, for example the normalized weighting (Hirano and Imbens 2001, Millimet and Tchernis 2009). Abadie and Imbens (2011) develop the bias-correction method for matching estimators.

IPW Origins

- Where do those weights come from? One can prove the following identities

$$ATE = \mathbb{E}\left[Y\frac{D-p(S)}{p(S)(1-p(S))}\right]$$
$$ATT = \frac{1}{\mathbb{P}(D=1)}\mathbb{E}\left[Y\frac{D-p(S)}{1-p(S)}\right]$$
$$ATU = \frac{1}{\mathbb{P}(D=0)}\mathbb{E}\left[Y\frac{1-p(S)}{D-(1-p(S))}\right]$$

Math Time: Proof for ATT

- Consider the following term

$$\begin{aligned}
\mathbb{E}\left[Y\frac{D-p(S)}{1-p(S)} \mid S\right] &= \mathbb{E}\left[Y\frac{D-p(S)}{1-p(S)} \mid S, D=1\right]\mathbb{P}(D=1 \mid S) \\
&\quad + \mathbb{E}\left[Y\frac{D-p(S)}{1-p(S)} \mid S, D=0\right]\mathbb{P}(D=0 \mid S) \\
&= \mathbb{E}[Y \mid S, D=1]p(S) \\
&\quad - \frac{p(S)}{1-p(S)}\mathbb{E}[Y \mid S, D=0](1-p(S)) \\
&= p(S)\left(\mathbb{E}[Y \mid D=1, S] - \mathbb{E}[Y \mid D=0, S]\right) \\
&= p(S)ATT(S)
\end{aligned}$$

Math Time: Proof for ATT

- Now we have the following expression for the conditional ATT

$$ATT(S) = \frac{\mathbb{E}\left[Y \frac{D-p(S)}{1-p(S)} \mid S\right]}{\mathbb{P}(D=1 \mid S)}$$

Math Time: Proof for ATT

- The unconditional ATT obtains after taking the expectation:

$$
\begin{aligned}
ATT &= \mathbb{E}_{S|D=1} ATT(S) \\
&= \sum_{k=1}^{\bar{S}} \frac{\mathbb{E}\left[Y \frac{D-p(S)}{1-p(S)} \mid S=k \right]}{\mathbb{P}(D=1 \mid S=k)} \mathbb{P}(S=k \mid D=1) \\
&= \sum_{k=1}^{\bar{S}} \mathbb{E}\left[Y \frac{D-p(S)}{1-p(S)} \mid S=k \right] \frac{\mathbb{P}(S=k)}{\mathbb{P}(D=1)} \\
&= \frac{1}{\mathbb{P}(D=1)} \sum_{k=1}^{\bar{S}} \mathbb{E}\left[Y \frac{D-p(S)}{1-p(S)} \mid S=k \right] \mathbb{P}(S=k) \\
&= \frac{1}{\mathbb{P}(D=1)} \mathbb{E}\left[Y \frac{D-p(S)}{1-p(S)} \right]
\end{aligned}
$$

ATT Weights

- The sample equivalent of the ATT formula is

$$\begin{aligned}\widehat{ATT} &= \frac{n}{n^1}\frac{1}{n}\sum_{k=1}^{n} y_k \frac{d_k - p_k}{1-p_k} \\ &= \frac{1}{n^1}\left(\sum_{i\in I} y_i - \sum_{j\in J}\frac{p_j}{1-p_j}y_j\right) \\ &= \frac{1}{n^1}\sum_{i\in I}\left(y_i - \sum_{j\in J}\frac{1}{n^1}\frac{p_j}{1-p_j}y_j\right)\end{aligned}$$

- The weights on the control units are

$$w_{ij} = \frac{1}{n^1}\frac{p_j}{1-p_j}$$

Homework

1. Prove the formulas for ATU and ATE
2. Using the sample equivalents of the formulas for ATU and ATE, derive the weights and show that they are equal to the ones we defined above.

Next Time on *Impact Evaluation Methods*...

Regression