

2. Matching estimators

LPO 8852: Regression II

Sean P. Corcoran

Selection bias

Lecture 1 showed why the simple difference in means between the treated and untreated cases does not provide the ATT:

$$\begin{aligned} E(Y|D=1) - E(Y|D=0) &= \\ E[Y(1)|D=1] - E[Y(0)|D=0] &= ATT + \underbrace{E[Y(0)|D=1] - E[Y(0)|D=0]}_{\text{selection bias}} \end{aligned}$$

Selection bias reflects baseline differences in $Y(0)$ between the $D=1$ and $D=0$ groups.

- Randomization of D would help!
- Regression can help under very strong conditions.

Matching

Matching estimators construct comparison groups that are *similar* according to a set of *matching variables*:

- Selecting specific matches
- Constructing a matched weighted sample
- Subclassification

The assumption: once we have conditioned on these matching variables—by selecting matches, constructing weights, or stratifying—treatment assignment and potential outcomes are independent. (The conditional independence assumption).

Weighted average

What is a weighted average? Given a weight for each observation i , the weighted average for Y is:

$$\frac{\sum_{i=1}^n w_i Y_i}{\sum_{i=1}^n w_i}$$

Weights are used for lots of reasons (Solon, Haider, & Wooldridge, 2015). In matching we may choose weights based on the values of confounders (i.e., matching variables).

Example 1: re-weighting

Imagine a job training program that serves 100 people where the outcome of interest (Y) is employment.

	Treated ($D_i = 1$)	Untreated ($D_i = 0$)	Diff (Mean Y)
Men	60 $Y = 1$ 20 $Y = 0$	350 $Y = 1$ 150 $Y = 0$	0.05
Women	12 $Y = 1$ 8 $Y = 0$	275 $Y = 1$ 225 $Y = 0$	0.05
Total	100	1000	
Mean(Y)	0.720	0.625	0.095
Mean(Male)	0.800	0.500	

Source: Huntington-Klein ch. 14

Example 1: re-weighting

The simple difference in means is: $E(Y|D = 1) - E(Y|D = 0) = \mathbf{0.095}$. However, you'll notice that for both men and women the treatment effect is only 0.05.

Assuming that *conditional on gender*, D is independent of potential outcomes, then each gender group is like a "randomized trial." The true treatment effect is **0.05** (for both men and women).

The simple difference in means *overstates* the treatment effect because the treatment group is disproportionately male, and males have a higher $Y(0)$ in the untreated state. There is selection bias (of 0.045).

Example 1: re-weighting

Can we re-weight the untreated group so that it “looks like” the treated group?

	Treated	Untreated	Untreated weight
Men	80	500	0.16 (80/500)
Women	20	500	0.04 (20/500)

We want the 500 untreated men to represent the 80 in the treatment group, so each gets a weight of 0.16. We want the 500 untreated women to represent 20 in the treatment group, so each gets a weight of 0.04. Note the weights sum to $(0.16 * 500) + (0.04 * 500) = 100$

Example 1: re-weighting

Using the weights, what is the proportion male in the untreated group? The proportion employed (Y)?

$$E_w(\text{male} | D = 0) = ((500 * 1 * 0.16) + (500 * 0 * 0.04)) / 100 = 0.80$$

$$E_w(Y | D = 0) = ((350 * 1 * 0.16) + (275 * 1 * 0.04)) / 100 = 0.67$$

Use this re-weighted mean to get the ATT:

$$ATT = 0.72 - 0.67 = 0.05$$

Note with the weights, the two samples are **balanced** on gender.

Example 1: re-weighting

To re-iterate:

- Gender was the only confounding factor here. Conditional on gender, treatment assignment was “as good as random.”
- We adjusted for differences in gender between the two groups using weights.
- The weights were chosen based on the distribution of gender in the treatment group (for ATT).
- We could also choose weights based on the distribution of gender *overall* for ATE

See the do-file *Lecture 2 weighting example* for this example in Stata.

Example 1b: subclassification

In this example we could alternatively use **subclassification**: grouping treated and untreated observations into strata, calculating differences within strata, and then weighting those differences to get a treatment effect estimate. The weights here are chosen based on the full sample (for ATE):

$$ATE = \underbrace{(0.75 - 0.70) * (580/1100)}_{\text{men}} + \underbrace{(0.60 - 0.55) * (520/1000)}_{\text{women}}$$

$$ATE = 0.05$$

Example 2: private vs. public colleges

Private				Public			
	Ivy	Leafy	Smart	All State	Tall State	Altered State	Earnings
A	1	Reject	Admit		Admit		110000
	2	Reject	Admit		Admit		100000
	3	Reject	Admit		Admit		110000
B	4	Admit		Admit		Admit	60000
	5	Admit		Admit		Admit	30000
C	6		Admit				115000
	7		Admit				75000
D	8	Reject		Admit	Admit		90000
	9	Reject		Admit	Admit		60000

Source: Angrist & Pischke *MM* (2015). Shaded cell represents the student's chosen college, from those they were admitted to. Based on Dale & Krueger (2002).

Example 2: private vs. public colleges

In the above table:

$$E[Y(1)|D = 1] - E[Y(0)|D = 0] = 92,000 - 72,500 = 19,500$$

$$= ATT + \underbrace{E[Y(0)|D = 1] - E[Y(0)|D = 0]}_{\text{selection bias}}$$

It is likely the treated group has a higher $Y(0)$ than the untreated group. This is suggested above by the higher mean earnings for students who applied and were admitted to private colleges (esp. groups A and C).

Example 2: private vs. public colleges

What if we could create equivalent groups by conditioning on some X ?
For example, what if:

$$\underbrace{E[Y(0)|D=1, X]}_{\text{unobserved}} = \underbrace{E[Y(0)|D=0, X]}_{\text{observed!}}$$

In other words, there is no difference in potential outcomes $Y(0)$ between $D=0$ and $D=1$, once we condition on X . Then we could contrast the mean Y for each set of X and then average them.

In the private vs. public college example, assume there is no difference in $Y(0)$ conditional on application/admitted group A-D:

Example 2: private vs. public colleges

	Ivy	Leafy	Smart	All State	Tall State	Altered State	Earnings
A	1	R	A		A		110000
	2	R	A		A		100000
	3	R	A		A		110000
B	4	A		A		A	60000
	5	A		A		A	30000
C	6	A					115000
	7	A					75000
D	8	R		A	A		90000
	9	R		A	A		60000

$Avg(Y|D=1, \text{Group}=A)=105,000$

$Avg(Y|D=0, \text{Group}=A)=110,000$. Difference = $105,000 - 110,000 = -5,000$

$Avg(Y|D=1, \text{Group}=B)=60,000$

$Avg(Y|D=0, \text{Group}=B)=30,000$. Difference = $60,000 - 30,000 = 30,000$

Example 2: private vs. public colleges

The simple average of the within-group differences (groups A and B) is:

$$(-5,000 + 30,000)/2 = \$12,500$$

A *weighted* average gives more weight to the group with more individuals:

$$(-5,000) * (3/5) + (30,000) * (2/5) = \$9,000$$

The weighted average uses the data more efficiently, and also generalizes appropriately to the groups included in the calculation. Note groups C and D are either all treated (private college) or all untreated (public college). There is no **common support** here. This term will come up again.

Example 2: private vs. public colleges

Note in this case that neither the weighted nor unweighted average estimates the ATE or ATT. This is due to the lack of common support.

- Without a counterfactual for the treated in group C, we can't estimate ATT (or ATE)
- Without a counterfactual for the untreated in group D, we can't estimate ATU (or ATE)

An illustration of the importance of being attentive to the population to which you are able to generalize with the data you have.

Example 2: private vs. public colleges

Angrist & Pischke *MM* (2015) explain how regression estimates are weighted averages of multiple matched comparisons. E.g., consider the regression:

$$Y_i = \alpha + \beta P_i + \gamma A_i + e_i$$

where $P_i = 1$ if the student attended a private college and $A_i = 1$ if the student was in group A (versus B). Students in groups C and D are excluded.

Using the Example 2 data, $\hat{\beta} = 10,000$. This is comparable to the averages on the previous slide, but not identical to either. Regression effectively applies different weights, but the idea is the same. (See *MM* for details).

We will return later to the differences between matching and regression.

Example 3: Catholic schools

Murnane & Willett (ch. 12) stratify the NELS sample by family income to estimate the effect of Catholic high school attendance on 12th grade math achievement:

Table 12.1 Descriptive statistics on annual family income, by stratum, overall and by type of high school attended, and average twelfth-grade mathematics achievement by income stratum and by high-school type (n = 5,671)

Stratum	Average Base-Year Annual Family Income (1988 dollars, 15-point ordinal scale)	Cell Frequencies	Average Mathematics Achievement (12th grade)			
			Public Catholic		Diff.	
Label	Income Range	Sample Variance	Sample Mean	Public Catholic	Public Catholic (% of stratum total)	
<i>Ht_Inc</i>	\$35,000 to \$74,999	0.24	11.38	11.42	1,969 344 (14.87%)	53.60 55.72 2.12***†
<i>Med_Inc</i>	\$20,000 to \$34,999	0.22	9.65	9.73	1,745 177 (9.21%)	50.34 53.86 3.52***†
<i>Lo_Inc</i>	≤\$19,999	3.06	6.33	6.77	1,365 71 (4.94%)	46.77 50.54 3.76***†
						Weighted Average ATE 3.01
						Weighted Average ATT 2.74

†p < 0.10; *p < 0.05; **p < 0.01; ***p < 0.001

†One-sided test.

Example 3: Catholic schools

Take the difference within each strata and then take the weighted average of these differences across strata.

The ATE uses *total* cell sizes as weights; ATT uses counts of *treated* cases in each cell as weights. These are smaller than the unconditional mean differences in math scores ($\hat{\beta}_{CATH} = 3.895$), suggesting upward bias.

Note income is a continuous variable. M&W created three strata with the aim of (1) creating balance in family income within each strata; (2) maintaining common support.

Again, we are appealing to the conditional independence assumption. Conditional on income (strata), enrollment in Catholic school is “as good as random” (!).

Return of the “unobservables”



Example 3: Catholic schools

Can also stratify on multiple covariates, as M&W do here with income and a measure of prior achievement (12 total cells):

Table 12.2 Sample frequencies and average twelfth-grade mathematics achievement, by high-school type, within 12 strata defined by the crossing of stratified versions of base-year annual family income and mathematics achievement ($n = 5,671$)

Stratum		Cell Frequencies		Average Mathematics Achievement (12th Grade)		
Base-Year Family Income	Base-Year Mathematics Achievement	Public	Catholic	Public	Catholic	Diff.
<i>Hi_Inc</i>	<i>Hi_Ach</i>	1,159	227	58.93	59.66	0.72
	<i>MHi_Ach</i>	432	73	49.18	50.71	1.53 ^{*,†}
	<i>MLo_Ach</i>	321	38	42.75	44.23	1.48
	<i>Lo_Ach</i>	57	6	39.79	40.40	0.62
<i>Med_Inc</i>	<i>Hi_Ach</i>	790	93	57.42	59.42	2.00 ^{***,†}
	<i>MHi_Ach</i>	469	49	47.95	50.14	2.19 ^{***,†}
	<i>MLo_Ach</i>	390	33	41.92	44.56	2.64 ^{***,†}
	<i>Lo_Ach</i>	96	2	37.94	39.77	1.83
<i>Lo_Inc</i>	<i>Hi_Ach</i>	405	36	56.12	56.59	0.47
	<i>MHi_Ach</i>	385	13	47.12	48.65	1.53
	<i>MLo_Ach</i>	433	21	40.99	41.70	0.71
	<i>Lo_Ach</i>	142	1	36.81	42.57	5.76
				Weighted Average ATE		1.50
				Weighted Average ATT		1.31

^{*} $p < 0.10$; ^{*} $p < 0.05$; ^{**} $p < 0.01$; ^{***} $p < 0.001$

[†]One-sided test.

Curse of dimensionality

Finer strata may provide a stronger argument for the conditional independence assumption that treatment group membership is unrelated to potential outcomes (within strata), but they make it more and more difficult to achieve common support—the **curse of dimensionality**.

Matching on a single variable

Examples 1-3 all created balance on a single variable (gender, sets of colleges, income). There are *lots* of ways to do this. When matching, there are a lot of choices to make:

- 1 What will the matching criteria be?
- 2 Will you *select* matches, or use weights to create a matched sample?
- 3 If selecting matches, how many?
- 4 If constructing a matched weighted sample, how will weights decay with distance?
- 5 What is the *worst* acceptable match?

Treatment effect estimation is usually the easy part. The hard part is the above.

What will the matching criteria be?

The goal is to construct comparison groups that are “similar” on matching variables. What does “similar” mean?

- Exact matching
- Coarsened exact matching
- Distance matching (e.g. nearest neighbor)
- Propensity score matching (observations with similar *propensity* to be treated)

Select matches or use weights?

Selecting matches:

- Literally picking observations to be “in” or “out” based on some criteria.
- Usually if an observation is “in” it gets equal weight.
- Intuitively appealing and avoids situation where some observations get very large weights.

Constructing a matched weighted sample:

- Determine how close untreated observations are to treated observations.
- Weight based on similarity, or to make matched sample “look like” treated group.
- Has nice statistical properties and is less sensitive/noisy.

If selecting matches, how many?

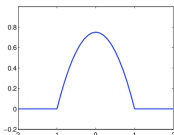
- One nearest neighbor? Or k nearest neighbors? Or **radius** matching (all neighbors within a given radius)?
- With replacement or without?

There is typically a **bias-variance tradeoff** in these decisions. More matches = larger sample size = less sampling variation. But more matches typically means “worse” matches, so more opportunity for bias.

How will weights decay with distance?

Typically a distance measure or a propensity score is used to construct a weight. We often want “less similar” observations to receive less weight.

A **kernel function** can do this. For example, the Epanechnikov kernel is $K(x) = \frac{3}{4}(1 - x^2)$ for $-1 \leq x \leq 1$ and 0 otherwise:



where x is a standardized distance measure. Note the weight is largest when the distance is 0 and then decays as you move away from 0.

How will weights decay with distance?

Propensity scores are often used to construct **inverse probability weights** where each treated observation gets a weight of 1 divided by their probability of treatment (p_i) and each untreated observations gets a weight of 1 divided by their probability of no treatment ($1 - p_i$).

As we will see, IPW makes the treated and untreated groups more similar:

- Treated observations with the biggest weights ($1/p_i$) are those that are more like the *untreated* group.
- Untreated observations with the biggest weights ($1/(1 - p_i)$) are those that are more like the *treated* group.

What is the worse acceptable match?

- Can choose a **caliper** or **bandwidth** for acceptable matches (in terms of the distance measure, or propensity score).
- Note a kernel implies a bandwidth
- Exact matching requires exact matches (as the name implies!)
- Coarsened exact matching requires exact matches on the coarsened continuous variable(s).

Again this decision involves a bias-variance tradeoff. There are methods for choosing optimal bandwidths based on some criteria.

Matching on multiple variables

When matching on *multiple* variables, we have all of the same decisions above to make. But we will need to reduce multiple differences into one dimension. Common approaches:

- Euclidean distance $||X_i - X_j|| = \sqrt{\sum_{m=1}^k (X_{mi} - X_{mj})^2}$, though variables are on different scales
- *Normalized* Euclidean distance—scales each variable by its variance:
$$\sqrt{\sum_{m=1}^k \frac{(X_{mi} - X_{mj})^2}{\sigma_m^2}}$$
- **Mahalanobis distance**—accounts for any covariance between x 's
- **Propensity scores**

With multiple matching variables, we can even combine criteria, like exact matching for one or more variables and distance matching for the others.

Mahalanobis distance

Take two observations (1 & 2) with X vectors of values X_1 and X_2 . The Mahalanobis distance measure is:

$$d(X_1, X_2) = \sqrt{(X_1 - X_2)' C^{-1} (X_1 - X_2)}$$

Loosely, this is the sum of square distances between values in X_1 and X_2 divided by the covariance. (C is the covariance matrix for the matching variables in X). If there is no covariance between the X , this reduces to the normalized Euclidean distance. Why “take out” the covariance?

- Suppose there is some latent characteristic that shows up in multiple matching variables. If those multiple variables are used to calculate distance, we may be “double-counting” by using distance on all of those variables.

Propensity scores

Think of the **propensity score** as a “one-number summary” capturing the relationship between treatment and X : $P(X_i) = Pr(D_i = 1|X_i)$. It is the probability of treatment given X .

The propensity score is often estimated using a binary logistic model:

$$P(D_i|X_i) = \frac{1}{1 + e^{-X_i\beta}}$$

Taking the logit tranformation results in a linear function of X :

$$\log\left(\frac{P}{1-P}\right) = X_i\beta$$

Curse of dimensionality, revisited

The curse of dimensionality comes up again when trying to match on multiple variables. The more matching variables you have, the less likely it is you will find a “close” match. Getting a better match on one variable x_1 may entail a worse match on x_2 .

Selecting matches in practice

Let's see some examples of matching when our aim is to select specific matches (i.e., we are not just creating weights for the purpose of re-weighting).

NOTE: there are lots of methods and decision points. It is easy to get lost in the weeds. But the objective is ultimately the same:

- **Create balance so that you can appeal to the CIA!**

You also want common support. For example, if you are estimating the ATT, you want overlap between your treated group (their X 's, or propensity scores) and your untreated group.

Exact matching

As the name suggests, **exact matching** entails pairing each treated (untreated) observation with one or more untreated (treated) observations with the same X . Estimate the ATT with:

$$\widehat{ATT} = \frac{1}{N_T} \sum_{D_i=1} (Y_i - Y_{j(i)})$$

where $Y_{j(i)}$ represents the Y for the matched case(s) for treated observation i . If multiple exact matches are used, $Y_{j(i)}$ stands in for the average of these.

Approximate matching

Approximate matching or distance matching relaxes the demand for an exact match and identifies “nearest neighbors” based on one or more covariates.

- Euclidean distance
- *Normalized* Euclidean distance
- Mahalanobis distance

Stata's teffects nnmatch

Stata's `teffects` implements a wide array of treatment effect estimators using matching, weighting, regression adjustment, etc. `teffects nnmatch` uses exact or approximate matching, or a combination of these.

```
teffects nnmatch (y x) (t), options
```

Here y is the outcome, x are the covariates, and t is the treatment indicator. In the options can specify `ate` or `atet`, and `ematch(vars)` to specify a list of variables on which you desire an exact match. For nearest neighbor matching you can specify the distance metric used, e.g., `metric(euclidean)`. There are lots of other options.

See simple matching examples on Github using NHIS and simulated data.

Stata's teffects nnmatch

More on `nnmatch`:

- The default number of nearest neighbor matches is 1, but in the case of exact matching, `teffects` will use *all* available exact matches. Note for exact matching to work for ATT, there must be at least one exact match for every treated observation.
- Can include `nneighbor(#)` option to specify the number of neighbors (although this doesn't seem to work properly for me).
- Can choose a `caliper(#)` in the options to specify “how bad” the match can be.
- Can include the option `gen(stubname)` to have Stata create new variables with the observation numbers of nearest neighbor matches. Note sort order can change—you may wish to id your observations to preserve sort order.